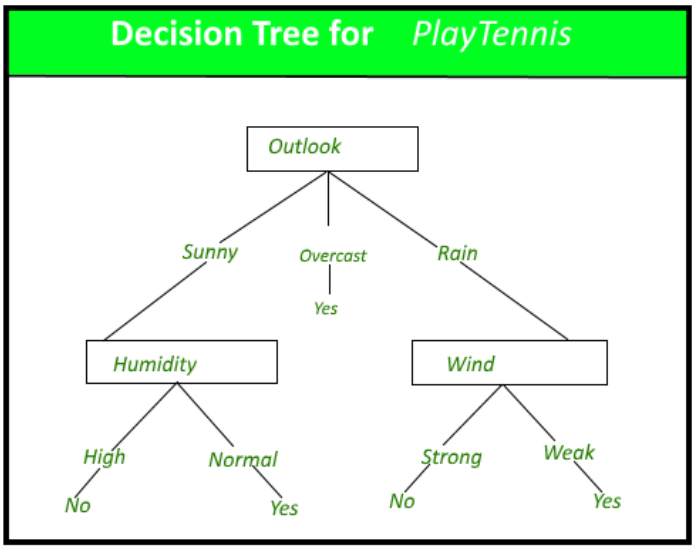
DECISION TREE REPORT

1. Describe the method that you use?

A Decision tree is used to visually and explicitly represent decisions and decision making. It is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
The Decision tree breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.
An example of the decision tree is shown in the below figure:



The advantages of decision tree methods are:

- Decision trees are able to generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

The disadvantages of decision tree methods are:

- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
- Decision tree can be computationally expensive to train.

In Decision Tree the major challenge is to identification of the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

1. Gini Index
2. Entropy

$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

$$Entropy = \sum_{i=1}^{n} -p(c_i)log_2(p(c_i))$$

where $p(c_i)$ is the probability/percentage of class c_i in a node.

Entropy is a measure of uncertainty associated with a random variable. The entropy increases with the increase in uncertainty or randomness and decreases with a decrease in uncertainty or randomness.
Gini impurity is a measure of misclassification, which applies in a multi-class classifier context.

2. Describe the dataset, and explain how you selected the two attributes?

The census dataset consists of 16 different attributes among which one of it is a class attribute which defines the target value for any given row in the dataset. The data set contains the following attributes and its corresponding datatypes.

Date	object
Age	int64
WorkClass	object
fnlwgt	int64
education	object
education-num	int64
marital-status	object
occupation	object
relationship	object
race	object
gender	object
capital-gain	int64
capital-loss	int64
hours-per-week	int64
native-country	object
class	object

Feature Selection is one of the core concepts in training the model which hugely impacts the performance of your model. The data features that you use to train your models have a huge influence on the performance you can achieve.

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

Benefits of performing feature selection before modelling your data?

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modelling accuracy improves.
- Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

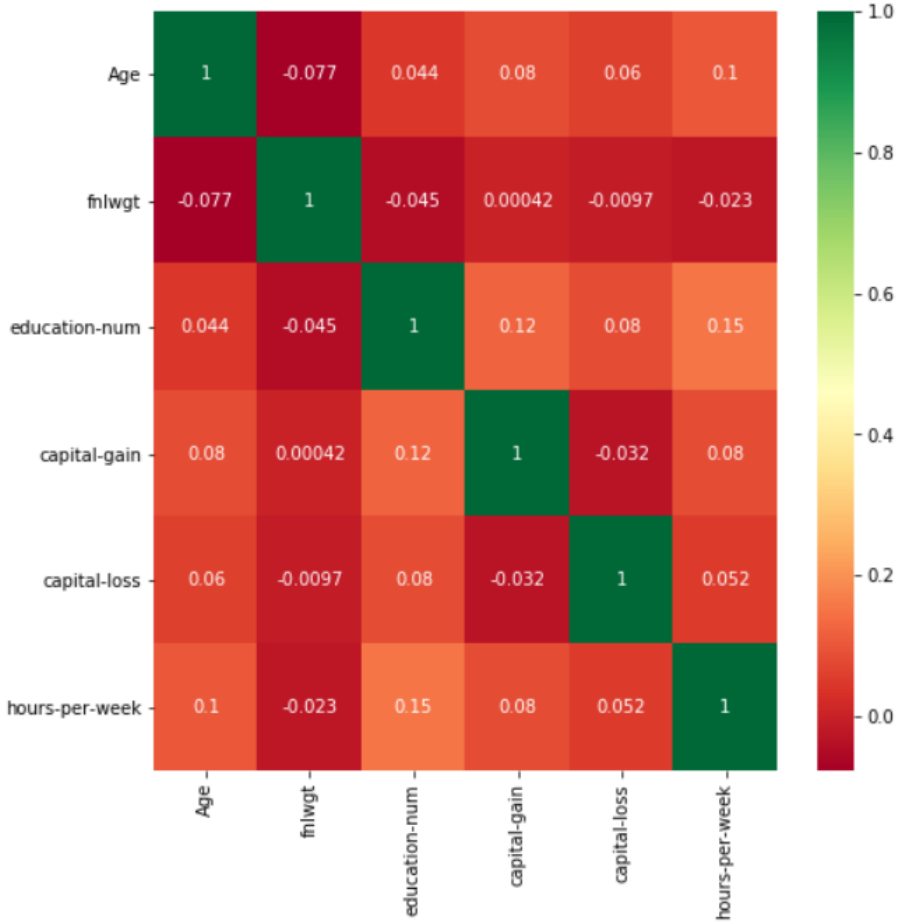
Here we do not consider the object datatypes as a feature to train the model as it would require for us to transform every object type to int or Boolean type which is understood by the model. Hence, we choose only the int datatypes.

We are going to select the features which play an important role in building the model. We are selecting the features by examining the co-relation between them. The features which have the highest co-relation with each other are selected to train the model. This method of selecting features is called Correlation Matrix with Heatmap.
Correlation states how the features are related to each other or the target variable. The value of the correlation can be:

1. Positive (increase in one value of feature increases the value of the target variable)
2. Negative (increase in one value of feature decreases the value of the target variable)

Heatmap makes it easy to identify which features are most related to the target variable, we will plot heatmap of correlated features using the seaborn library.

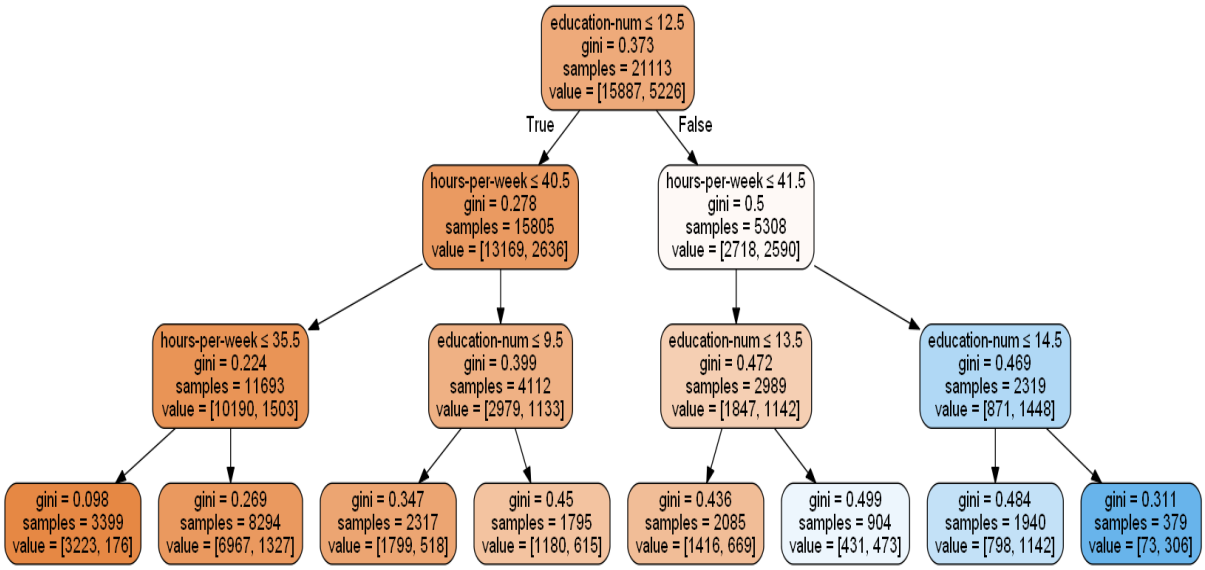
The following heatmap shows the correlation among the attributes in the census dataset:



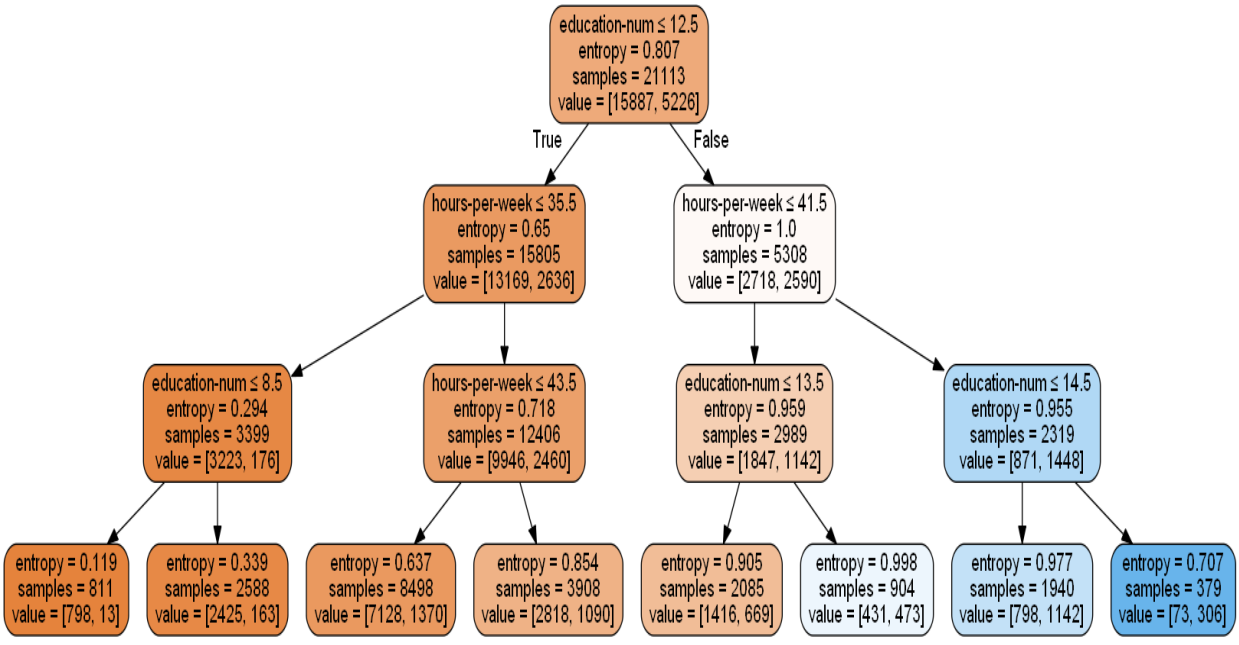
From the above correlation matrix, we can see that the education-num attribute and the hours-per-week attribute have the highest correlation when compared to other attributes in the dataset.
Hence, we will choose education-num and hours-per-week attribute in training our model so that we could attain maximum accuracy and precision.
Here we are giving two attributes to the decision tree which has the highest correlation so that we could derive rules and the target values easier than giving it all the attributes and lowering the accuracy and precision. Hence, we train our decision tree model based on the two attributes to increase the accuracy and precision.

3. Visualize the classifier in a 2D projection and compare gini and entropy.

Using Gini as a criterion for classification we get the following Decision tree:



Using Entropy as a criterion for classification we get the following Decision tree:



From the above Decision trees we can see that the split of the decision tree from the root upto level two is same for both gini and entropy. But from the level two we notice that there is a change in the split of the values. Also when we compare the precision and accuracy of both , it turns out to be the same, only difference is that Gini will find the largest class, and entropy tends to find groups of classes that make up ~50% of the data. Gini minimizes the misclassification.

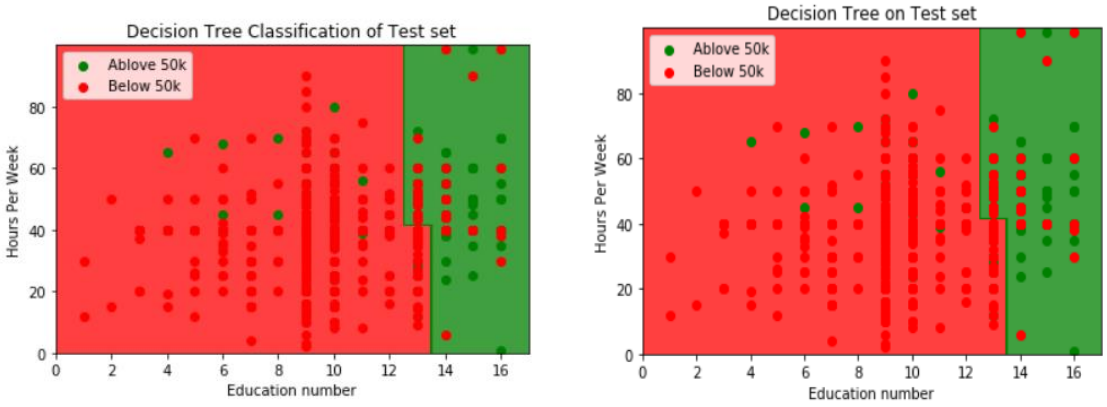
Following decisions can be concluded on the basis of the decision tree using gini:

1. If a test data has education-num <=12.50 and hours-per-week <=40.5 then it belongs to the class <=50K
2. If the test data has education-num >12.50 and hours-per-week<=41.5 then
   i) If education-num<13.5 then it belongs to class <=50K
   ii) If education-num>13.5 then it belongs to class >50K
3. If the test data has education-num>12.50 and hours-per-week <=14.5 and >14.5 then it belongs to class >50K

Following decisions can be concluded on the basis of the decision tree using entropy:

1. If a test data has education-num <=12.50 and hours-per-week <=35.5 then it belongs to the class <=50K
2. If the test data has education-num >12.50 and hours-per-week<=41.5 then
   i) If education-num<13.5 then it belongs to class <=50K
   ii) If education-num>13.5 then it belongs to class >50K
3. If the test data has education-num>12.50 and hours-per-week <=14.5 and >14.5 then it belongs to class >50K

Following is the plot of graphs for the values predicted by the classifier against actual values:



On the left we see the graph for gini and on the right we see the graph for entropy. From the above graphs we can understand which of the values are predicted right and wrong by the classifier.

REFERENCES:

[1]https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b

[2]https://stackoverflow.com/questions/23543406/getting-feature-names-in-addition-to-values-scikitlearnpandas

[3]https://stackoverflow.com/questions/18438997/why-is-pydot-unable-to-find-graphvizs-executables-in-windows-8

[4]https://datascience.stackexchange.com/questions/37428/graphviz-not-working-when-imported-inside-pydotplus-graphvizs-executables-not

[5]https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

[6]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html#sklearn.metrics.classification_report

[7]https://stackoverflow.com/questions/19233771/sklearn-plot-confusion-matrix-with-labels
[8]https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
[9]https://en.m.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity
[10]https://www.quora.com/What-is-difference-between-Gini-Impurity-and-Entropy-in-Decision-Tree
[11]https://www.geeksforgeeks.org/decision-tree-introduction-example/
[12]https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1
[13]https://www.saedsayad.com/decision_tree.htm
[14]https://scikit-learn.org/stable/modules/tree.html
[15]https://scikit-learn.org/stable/auto_examples/tree/plot_iris_dtc.html
[16]https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1