Semester Report Assignment

Andrei Nicolaiciuc

2021.12.10

Abstract

In this project were implemented and compared classification (Decision Tree Classifier, Random Forest Classifier, Linear Regression Classifier, Logistic Regression Classifier, k-Nearest Neighbours Classifier), clustering (K-Means Clustering, Hierarchical Agglomerative Clustering) and Apriori Frequent Pattern Mining.

Keywords: Data Science, Classification, Clustering.

1 Introduction

The value of Data Science combines domain expertise from programming, mathematics, and statistics to generate insights and make sense of data. When someone considers why data science is becoming increasingly essential, the answer is that the value of data is skyrocketing.

Data Science helps businesses to easily comprehend massive amounts of data from a variety of sources and gain important insights to make better data-driven choices. Data science is in great demand since it explains how digital data is reshaping organizations and assisting them in making more informed and essential decisions.

By examining several pattern recognition techniques, this project seeks to discover the best model to predict if an individual's income is larger than 50k dollars using data from the UCI Machine Learning repository's Adult Data Set.

2 Data Set Information

The Adult Data Set contains 32561 entries and 15 columns. Each entry contains the following information about an individual: age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income.

Extraction was done by Barry Becker from the 1994 Census database. By studying various pattern recognition approaches, this project attempts to discover the best model to predict if an individual's salary is larger than 50k based on data supplied by the UCI Machine learning repository dubbed the 'Adult data set.'

3 Data Preprocessing

Machine learning algorithms learn a mapping from input variables to a target variable during a predictive modeling assignment. The most frequent type of predictive modeling project utilizes structured data, often known as tabular data. This is data in the form of a spreadsheet or a matrix, with rows of samples and columns of characteristics for each one.

Data scantiest can't fit and assess machine learning algorithms on raw data; instead, data scantiest have to alter the data to satisfy the needs of particular machine learning algorithms. Furthermore, in order to get the highest performance given our available resources on a predictive modeling project, data scantiest must select a data representation that optimally exposes the unknown underlying structure of the prediction issue to the learning algorithms.

Data integration is the process of combining data from several sources into a unified data storage, such as a data warehouse. Data reduction can reduce data size by aggregating, deleting duplicate characteristics, or clustering, for example.

Pre-processing was carried out in order to make the provided raw data acceptable for classification. The characteristics belonged to one of the following categories:

Numerical Features (6 in total): These features give information in numerical format. This category includes age, fnlwgt, education number, capital gain, capital loss, and hours per week.

Categorical Features (8 in total): These features give information in a

categorized manner. Work class, education, marital status, occupation, relationship, race, gender, and home country are all characteristics that fall under this group.
Handling missing values: Some attributes (which are merely categorical in this case) for certain data points have unknown values. These unknown values must be replaced with the value that best matches the situation.
Removing the rows containing missing values: In some ways, this is the simplest and best strategy for dealing with missing data because the whole categorization is based on the actual data set, with no assumptions made about the missing values, and hence the outputs will be a real picture of the obtained input data.

Also, the information was pre-processed in order to be utilized for Frequent Pattern Mining. As can be observed, the data contains a few meaningless properties, such as 'fnlwgt' and 'education-num'. These characteristics were removed from the data set. After removing unnecessary attributes, the total number of attributes is 13. It also features a question mark to signify missing values. As a result, rows with missing values were removed. After deleting missing values, the number of rows is 30162. Certain qualities contain continuous numeric data and must be quantized into bins before the technique can be used.

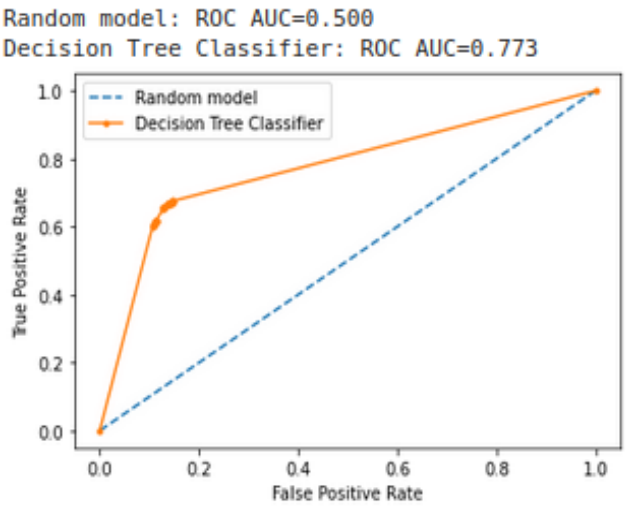
4 Classification

Classification is the process of categorizing a given collection of data into classes. It may be done with both organized and unstructured data. Predicting the class of provided data points is the first step in the procedure. The classes are also known as the goal, label, or categories.Approximating the mapping function from discrete input variables to discrete output variables is the problem of classification predictive modeling. The primary purpose is to determine which class/category the new data will belong to.

Classification is a supervised Machine Learning approach that predicts the class of incoming data based on previously classified data. Classification is a type of data analysis in which models defining relevant data classes are extracted. Classifiers are models that predict categorical (discrete, unordered) class labels. In the case of this data set, the following classifiers were used: Decision Tree Classifier, Random Forest Classifier, Linear Regression Classifier, Logistic Regression Classifier, k-Nearest Neighbors Classifier.

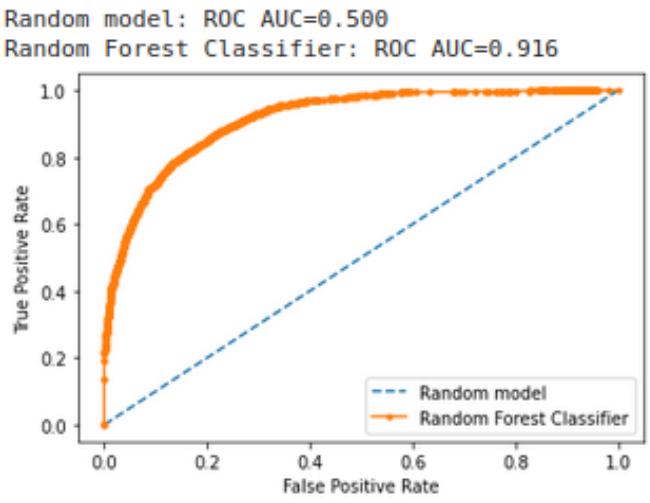
4.1 Decision Tree Classifier

This is a supervised learning approach with an implementation that mimics a tree structure. It starts with all training samples as a root node at the top and evaluates the feature that gives the most information gain at each subsequent layer. By repeating this procedure, it creates several subtrees over a subset of training samples categorized along the route. Scikit learn-tree module's DecisionTreeClassifier() function was utilized.



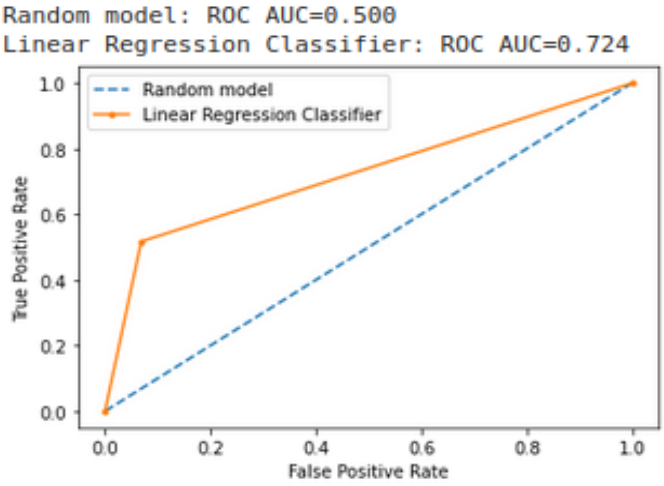
4.2 Random Forest Classifier

This is a popular supervised learning method since it produces excellent results without the need for hyperparameter adjustment most of the time. The basic concept here is that the random forest is a collection of several decision trees that combine their results to provide a more stable and accurate forecast. Unlike decision trees, which utilize the most representative feature to divide the node, the random forest takes the best feature from a random selection of characteristics to split the node. One noteworthy virtue of a random forest is that, when sufficient decision trees are utilized, it does not overfit the data in most circumstances. Scikit learn-ensemble model's RandomForestClassifier() function was utilized.



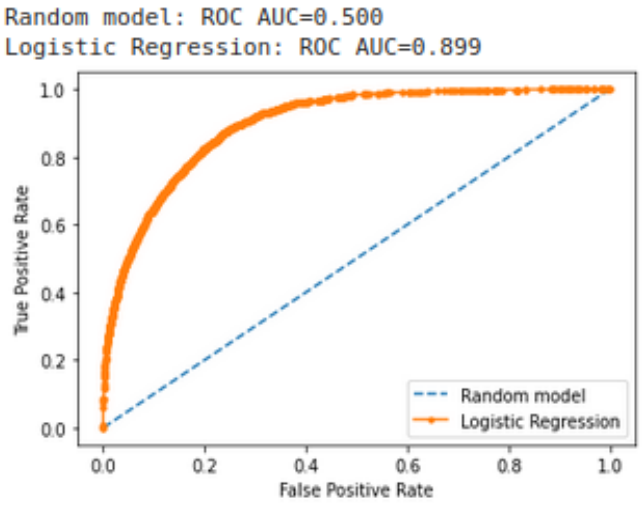
4.3 Linear Regression Classifier

Linear Regression is a supervised learning-based machine learning technique. It carries out a regression job. Based on independent variables, regression models a goal prediction value. It is mostly used to determine the link between variables and predicting. Different regression models differ in terms of the type of connection they evaluate between dependent and independent variables, as well as the number of independent variables utilized.



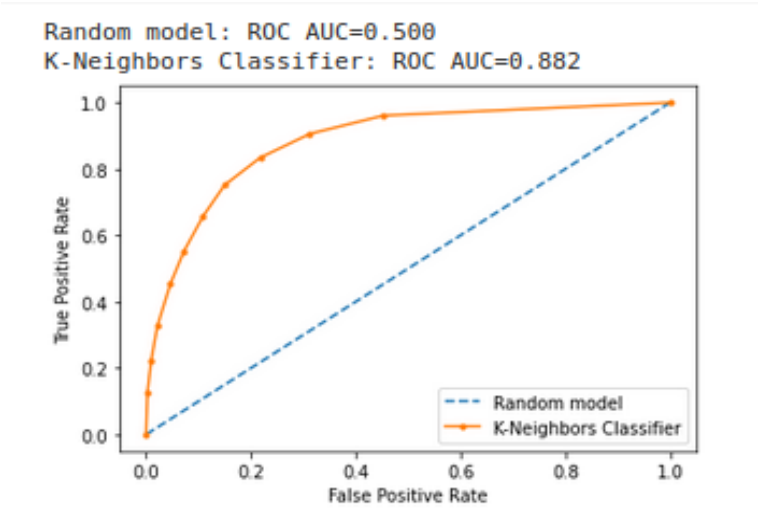
4.4 Logistic Regression Classifier

This type of statistical analysis (also known as logit model) is often used for predictive analytics and modeling, and extends to applications in machine learning. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite options A, B, C or D (multinomial regression). It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.



4.5 k-Nearest Neighbors Classifier

This is a non-parametric, supervised learning technique that classifies a given point based on its neighbors. Because the data point is assigned to the class of the nearest 'k' neighbors, the choice of the 'k' becomes critical. In most cases, the Euclidean norm is used as a distance metric to determine how near a data point is to the test data point. Once here exists identified such 'k' nearest data points, the test data is labeled based on a majority vote from the class labels of the 'k' nearest data points. Scikit learn – neighbors module provides the function KNeighborsClassifier().



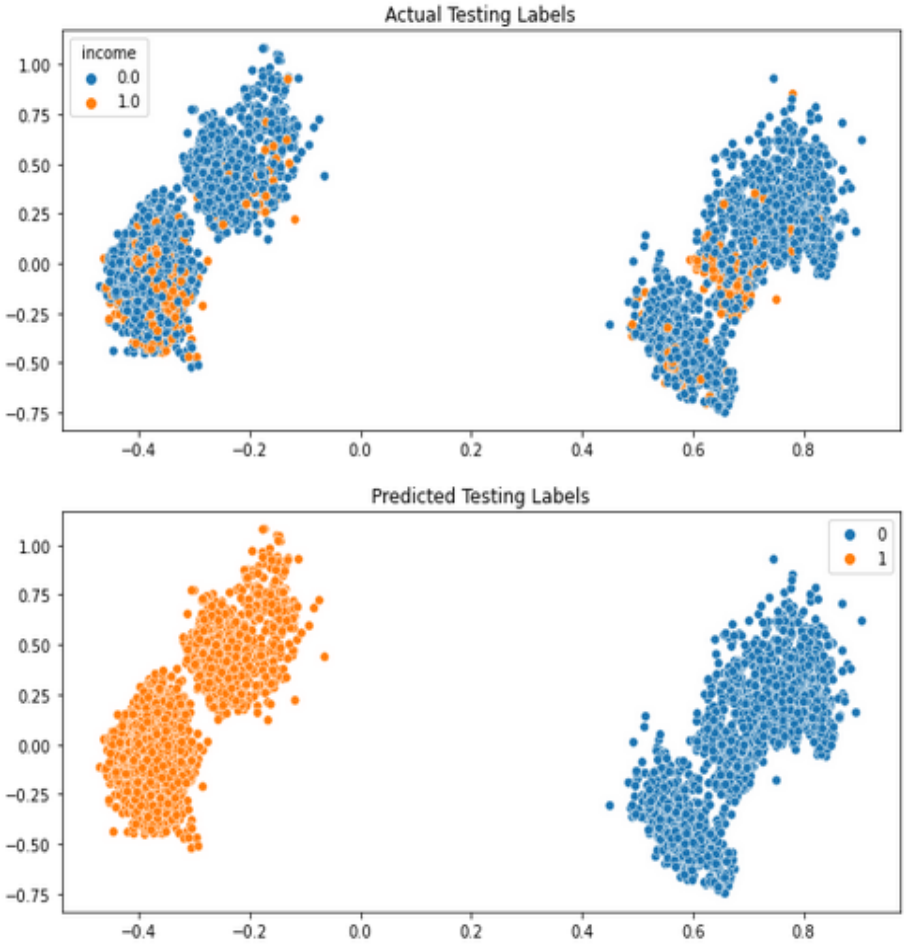
5 Clustering

Clustering is an unsupervised Machine Learning approach used to detect similarities between items in a collection. Cluster analysis, often known as clustering, is the process of dividing a collection of data items (or observations) into subsets. Each subset is a cluster, and things inside a cluster are comparable to one another but not to objects in other clusters. A clustering is the collection of clusters that emerges from a cluster analysis. Cluster analysis is widely utilized in a wide range of applications, including corporate intelligence, visual pattern recognition, Web search, biology, and security. For this data set, two separate techniques were used: KMeans and Hierarchical Agglomerative clustering. In a nutshell, the goal is to separate groups with similar characteristics and assign them to clusters.

5.1 K-Means Clustering

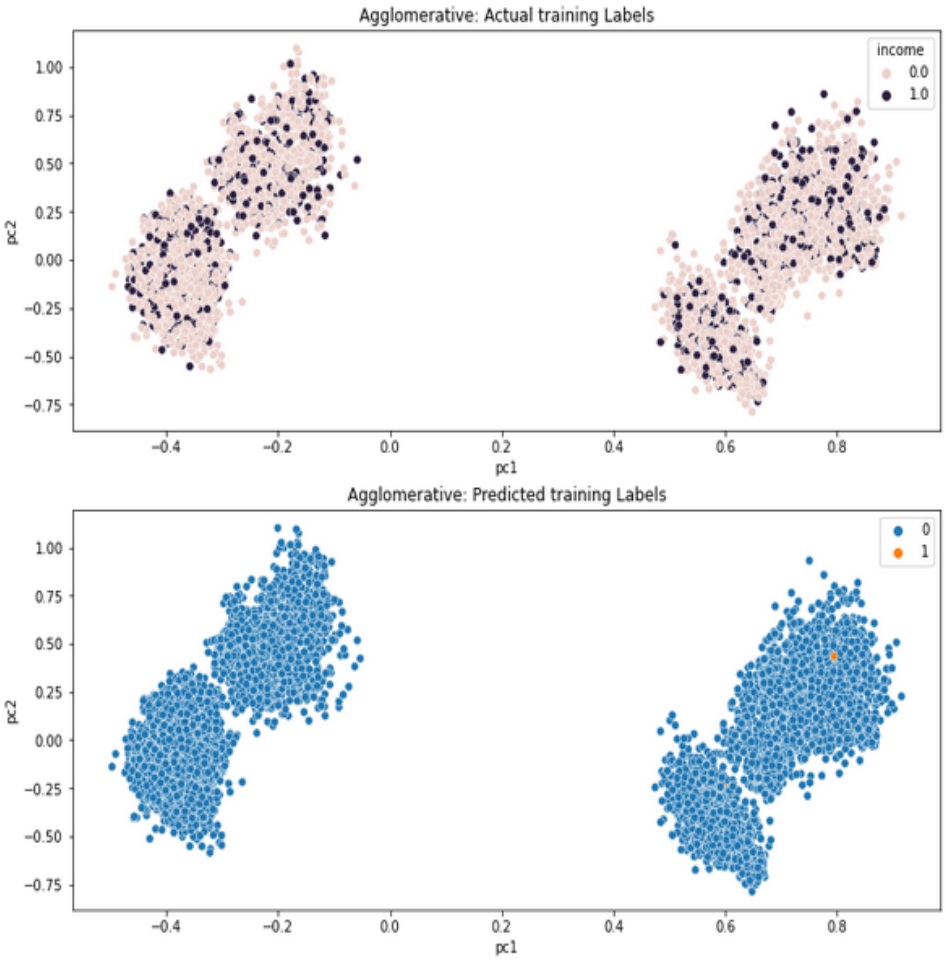
Because of its simplicity and efficacy, K-means is one of the most common unsupervised machine learning algorithms. Methods of partitioning: A partitioning method creates k partitions of data from a collection of n objects, where each partition represents a cluster. That is, it separates the data into k groups, each containing at least one item. In other words, partitioning algorithms partition data sets on a single level. The most common parti-

tioning methods use exclusive cluster separation. That is, each item 5must be assigned to just one group. This criterion can be eased using fuzzy partitioning approaches, for example. The bibliographic notes provide references to similar procedures. The majority of partitioning strategies are based on distance. A partitioning technique generates an initial partitioning given k, the number of partitions to produce. Before beginning the procedure, the essential step is to determine the K, number of clusters.



5.2 Hierarchical Agglomerative Clustering

The most popular sort of hierarchical clustering is agglomerative clustering, which is used to organize items into clusters based on their similarity. AGNES is another name for it (Agglomerative Nesting). The procedure begins by treating each item as if it were a singleton cluster. Following that, pairs of clusters are combined one by one until all clusters have been merged into one large cluster holding all items. The ultimate result is a dendrogram, which is a tree-based representation of the items.



6 Conclusion

Finally, utilizing Python and its strong libraries, was successfully developed a classification Machine Learning prediction model that forecasts whether a particular adult's income would be more than 50k or not.

The Random Forest classifier is found to perform well on adult data set. It has the best score - 0,916. This is based on the "Compare Classifiers". As a result, a Recall of 1.00 has been achieved from Agglomerative Classification and 0.61 from K-Means Classification. A Precision of 0.76 has been achieved from Agglomerative Classification and 0.69 from K-Means Classification. However, performing the right set of pre-processing steps and selecting the right set of features plays a significant role in the performance of the classifier.

Also was used the sampling improvement on the Apriori technique, which is based on computing Apriori Frequent Pattern Mining on a sample of the complete data set rather than repeatedly scanning the entire data set. In this manner, the complete data set is only scanned once, namely during the generation of the sample data set. Because the approach is applied to a smaller data set, the calculation time is dramatically decreased. With an optimal sample factor and support count, here can be produced all the worldwide frequent patterns, resulting in excellent efficiency and accuracy. However, most of the time when sampling improvement is utilized, accuracy is sacrificed in order to boost efficiency because the method is only applied to a portion of a larger data set. To improve accuracy, a lower support criterion is sometimes utilized instead of a minimum support threshold. This improvement is especially advantageous in applications where efficiency is crucial.