**Introduction**

The goal of this project is to classify US citizens into low income or high-income citizens based on the census data provided to us. This data contains attributes like Age, Gender, Education, Marital status etc. We consider this dataset and build a model to check how accurately we can classify the citizens into different income categories. This helps determine what features are affecting the income and what policies need changing.

**Data Cleansing**

Data cleansing is the process of identifying corrupt, inaccurate or irrelevant data. This is the most critical or important step in data science. Not cleaning the data properly could have dire consequences on the end results and the accuracy of the model. There are different ways in which we can clean the data:

1. We can remove the entire rows or columns
2. We can use different statistical methods to fill the missing data
3. Mean
4. Median
5. Mode
6. Imputation- infer them from the known part of the data

The steps followed for data cleansing is to first identify the critical data fields, collect the data, discard duplicate values, resolve empty values, standardize the cleaning process (which data is most often used, when it needs to be cleaned and who is responsible for the cleaning process). We then review, adapt, and repeat the process [1]

**Data Cleansing for Census Data**

In our dataset, we have 15 fields: age (39), work class (State-gov), final weight (77516), education (Bachelors), education num (13), marital status (Never Married), occupation (Adm- clerical), relationship (Not in family), race (white), sex (male), capital-gain (2174), capital-loss (0), hours per week (40), native country (united states) and salary (<=50k) [2]. The names in the brackets are those mentioned in the database given to us for the purpose of this assignment.

Chart

Description automatically generated with low confidenceFirst, we explore each of these features to check for any missing values. The dataset doesn’t contain any missing values. But work class and the native country columns have “?” as values which are not adding any value to the dataset. Hence, these rows are removed. Age, Education, Hours per week and gender are some of the most contributing features for the classification. The graphs are in Appendix A Graph 1,2,3 and 4.

These graphs indicate how much of a significant impact these features have on the prediction. In the final weight column, we see that the values greater than 450k are outliers and are having minimal impact. So, these rows are removed.

**Dependent and Independent Variables**

We now divide the columns into dependent and independent variables. We choose columns 1 to 14 as independent variables and the salary column as the dependent variable as our model is trying to predict and test this particular column based on the other features.

**Data Loading**

Initially we load the census data using scikit learn which is a machine learning library.

We then manually divide the data into training and testing sets. We use the training dataset to train the model and then we use the same model on our testing dataset so we can compare the results from training data and testing data to check the accuracy of the model. This is a good practice because if we use the training dataset to also test the model, we get the right results almost every time since the model is already trained using a particular data. But using a different set of data helps us understand the accuracy of categorization and validate the model.

**Nearest Neighbors**

A picture containing logo

Description automatically generatedChart, diagram

Description automatically generatedFor this classification, we use the K method also known as K nearest neighbor method. In this method, the K represents the number of neighbors we consider to determine the category of classification. This method of classification only considers numerical values so we convert our string values into numerical values using LabelEncoder from sklearn library. In this method, we consider K number of nearest neighbors to a particular test datapoint. We find these nearest neighbors using distance formula and Pythagorean theorem. Consider the image where d is the hypotenuse and the legs are the distances between the two coordinates. We then use the distance formula given below.

This is a lazy learning algorithm i.e.; it memorizes the training dataset. It is most suitable for categorizing. The highest number of categories out of all the K nearest neighbors is assigned to the test data. For example, we take K=3, so we consider 3 nearest neighbors out of which 2 are red and 1 is black. Our test point would then be classified as red.

We use multiple Ks to compare the results and check the accuracy. We compare these accuracies using graphs to see which k value has the highest accuracy. Please observe Appendix A graph 5 where we have plotted graph for k values from 1 to 14. The highest accuracy is 79 % and is for K=14. Refer to Appendix B for matrices.

**Confusion matrix and Classification report**

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. **The confusion matrix shows the ways in which your classification model** **is confused when it makes predictions [3]. The matrix predicts true positive(TP) in x[0][0] position, false positive(FP) in x[0][1] position, true negative(TN) in x[1][0] position and false negative(FN) in x[1][1] position.**

**A classification report has columns like precision, recall, f1 score and support. Precision is the percentage of predictions that were correct. For example for true positives the formula would be TP/(TP+FP). Recall is the percentage of positive cases that were caught. For example: TP/(TP+FN). F1 score is the percentage of positive predictions that were correct. Formula for F1 score would be 2\*(Recall\*Precision)/(Recall\* precision)[4] Support is the number of samples of the true response that lie in the class.**

**Conclusion**

From the above analysis we can conclude the following

1. The model can predict with 79% accuracy to which salary category one may belong to based on the features given in the dataset.
2. We can observe from the graph that increasing the K value has increased the accuracy of the model. From the graph, it is clear that we can pick K=14 for the best accuracy.

In conclusion, we can classify KNN as a very good accurate algorithm for classification of citizens into various salary categories. In our modern-day world, where we are trying to bridge the gender pay gap, where we are constantly trying to improve educational policies and reduce the illiteracy rate, where we are trying to invent something new every day, increase diversity, encourage investments, increase GDP and better living conditions for everybody, a model like this is very important. It can help us understand where we can improve and what policies to change to maximize development.

**Appendix B: Graphs**

Graph 1: Age vs Salary

Chart, histogram

Description automatically generated

Graph 2: Education vs Salary

Chart, bar chart

Description automatically generated

Graph 3: Hours per week vs Salary

Chart, bar chart

Description automatically generated

Chart, treemap chart

Description automatically generatedGraph 4: Gender vs Salary

Chart, line chart

Description automatically generatedGraph 5: K value accuracy

**Appendix B: Matrices**

Table

Description automatically generated with medium confidence

**References**

1. [What is Data Cleansing? Guide to Data Cleansing Tools, Services and Strategy | Talend](https://www.talend.com/resources/what-is-data-cleansing/)
2. [JuniorMkhatshwa/Demographic-Data-Analyzer (github.com)](https://github.com/JuniorMkhatshwa/Demographic-Data-Analyzer)
3. [**Jason Brownlee**](https://machinelearningmastery.com/author/jasonb/) on November 18, 2016 in [**Code Algorithms From Scratch**](https://machinelearningmastery.com/category/algorithms-from-scratch/)

[What is a Confusion Matrix in Machine Learning (machinelearningmastery.com)](https://machinelearningmastery.com/confusion-matrix-machine-learning/)

1. [Understanding the Classification report through sklearn – Muthukrishnan](https://muthu.co/understanding-the-classification-report-in-sklearn/)