Introduction:

Education is not just seen as a source of knowledge, but can be seen as a form of investment in human capital. Like financial assets, the purpose of investing in human capital is ultimately to make more money. The average salary with those of a high level of education (which includes at least one year of undergraduate studies) in the EU was approximately 50% higher than those with a medium level of education (highschool education) and a whopping 70% higher than that of people with low level of education according to Eurostat Data. However, with an increase in the proportion of people seeking higher education, the value of educational credentials have decreased, resulting in lower wages for highly educated workers.

To further investigate what kind of relationship exists between education level and income, this report looks at constructing a logistic regression model which aims at classifying the income of individuals based on their education level and other additional factors. Two models will be constructed using a Kaggle dataset based on the Census Income Data from the UCI Machine Learning Repository. One will be a binomial logistic regression where the independent and dependent variables are education level and income category respectively. The other will be also be a binomial logistic regression but will include other variables such as sex and hours worked per week. The models will be compared to see which is the better suited model for classification. The hypothesis is that the second model, the model with more variables, will perform better due to the assumption that it will account for confounding variables. This will serve to answer whether the education level variable alone is good enough to classify income categories or whether other additional variables need to be considered.

Dataset:

This report analyzes the Kaggle dataset based on the Census Income Data from the UCI Machine Learning Repository. The original cross sectional dataset, extracted from the 1994 Census database, contains 48842 instances with 14 different attributes. Categorical variables were removed, resulting in only 8 remaining variables. The income and sex variables were replaced with dummy variables. The summary statistics is given below:

Figure 1:

	age	fnlwgt	education.num	sex	capital.gain	capital.loss	hours.per.week	income
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	0.669205	1077.648844	87.303830	40.437456	0.240810
std	13.640433	1.055500e+05	2.572720	0.470506	7385.292085	402.960219	12.347429	0.427581
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	0.000000	40.000000	0.000000
50%	37.000000	1.783560e+05	10.000000	1.000000	0.000000	0.000000	40.000000	0.000000
75%	48.000000	2.370510e+05	12.000000	1.000000	0.000000	0.000000	45.000000	0.000000
max	90.000000	1.484705e+06	16.000000	1.000000	99999.000000	4356.000000	99.000000	1.000000

The definition of the variables in the dataset are as follows:
age: age of the individual
fnlwgt: weight of the number of people represented in the individual's working class in relation to the overall population
education-num: numerical representation of the education level
sex: sex of the individual (1: male , 0: female)
capital gain: money individual made from investments
capital loss: money individual lost from investments
hours per week: number of hours individual works during the week
income: category of income (1: >= 50k USD 0: <= 50k USD) *dependant variable

The data overall comprises individuals in their mid career based on the average age. The average education number value indicates that most individuals have had some college education. Around 40 hours tends to be the average hours worked per week which is close to the national average of 38.7 according to the labor force statistics for the US. An interesting observation not noted in the summary statistics above is that the dataset comprises 21790 men and 10771 women.

Methodology:

Logistic regression is a common machine learning algorithm, often used for binary classification. It is suitable for the current problem of classifying individuals' income into either above 50k or below 50k based on the independent variable(s).

To build a logistic regression model using the stated variables, sklearn library was utilized. Using sklearn, a logistic regression object was created, ready for fitting. The x values for the first model was the education-num variable. The x values for the second model were all the variables from the modified dataset above with the exception of the dependent variable, income. The y variable was income for both models. For future model evaluation, data was split into test and train, where 25% of the data was set aside for testing later. The x and y train data were then fitted with the logistic regression object, completing both models.

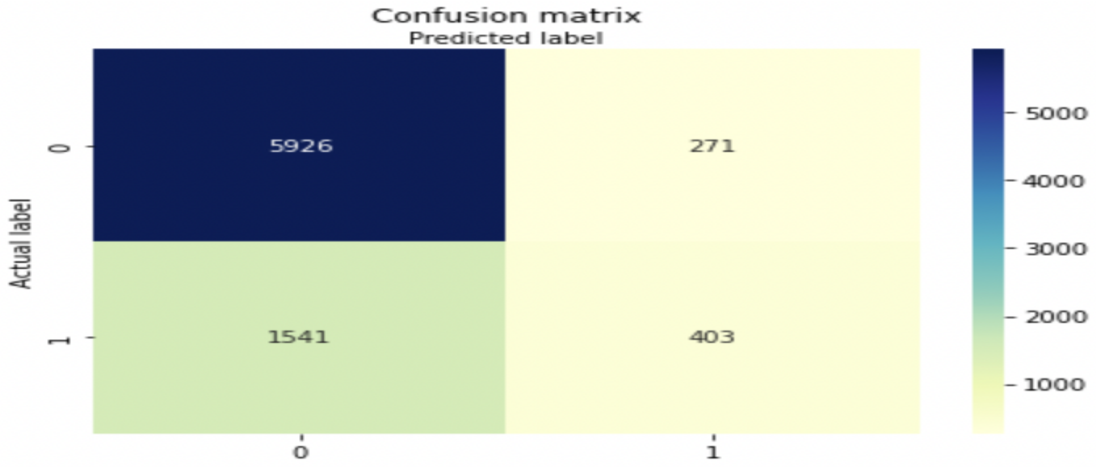
To evaluate the performance of both the classification models, confusion matrices were created using the x and y test data. A confusion matrix is a 2 by 2 matrix, where the diagonal values represent the accurate predictions while the non-diagonal values represent the inaccurate predictions. Mectrics class from sklearn was used to create the matrices for both models. To visualize the matrices, heatmap was created using the seaborn library. In addition to the heatmap, receiver operating characteristic curve (ROC) was also plotted for both models as an additional performance measurement. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR). TPR is defined as a true positive number divided by the sum of true positive and false negative. FPR is defined as the false positive divided by the sum of false positive and true negative. Once plotted, to calculate the performance from the plotted graph, the area under the curve (AUC) is calculated. An AUC score near 1 would indicate a perfect classifier while the AUC score near 0.5 would indicate a

terrible classifier. The AUC scores of both models will be calculated and compared to conclude which model is better.

Results:

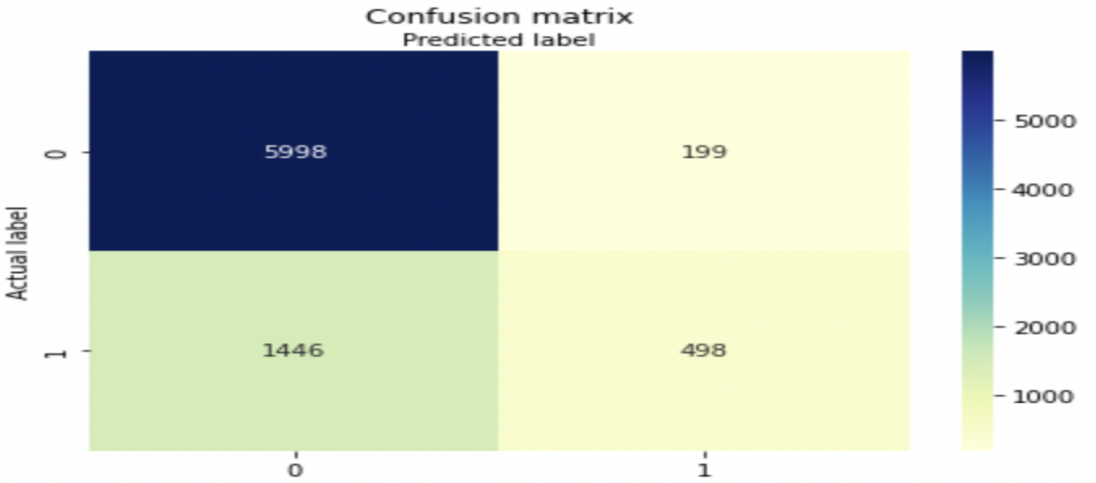
To visualize the matrices, heatmap was created using the seaborn library. Heatmaps for the models are given below:

Figure 2 (first model):



The heatmap above (figure 2) represents the confusion matrix for the 1st model which has education level as the only independent variable. The heat map below represents below (figure 3) represents the confusion matrix for the 2nd model which has several independent variables

Figure 3 (second model):



Based on the heatmaps above, it can be observed that the second model's True Negative (TN) and True Positive (TP) scores are higher than the first model's. In addition, it also has a lower False Positive (FP) and False Negative (FN) score. This would imply that the predictions made by the second model are more accurate than the first model, therefore, making it the better model. To reach such a conclusion, the true positive rate (TPR) was calculated using TP and FN while the false positive rate (FPR)was calculated using TN and FN for both models. The plots of the TPR against the FPR was created and provided below

Figure 4 (first model):

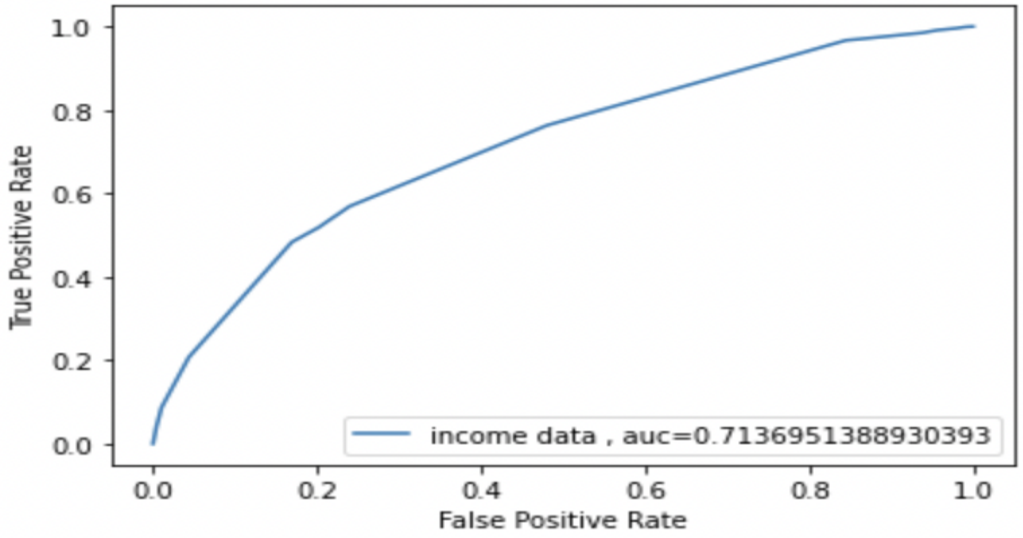
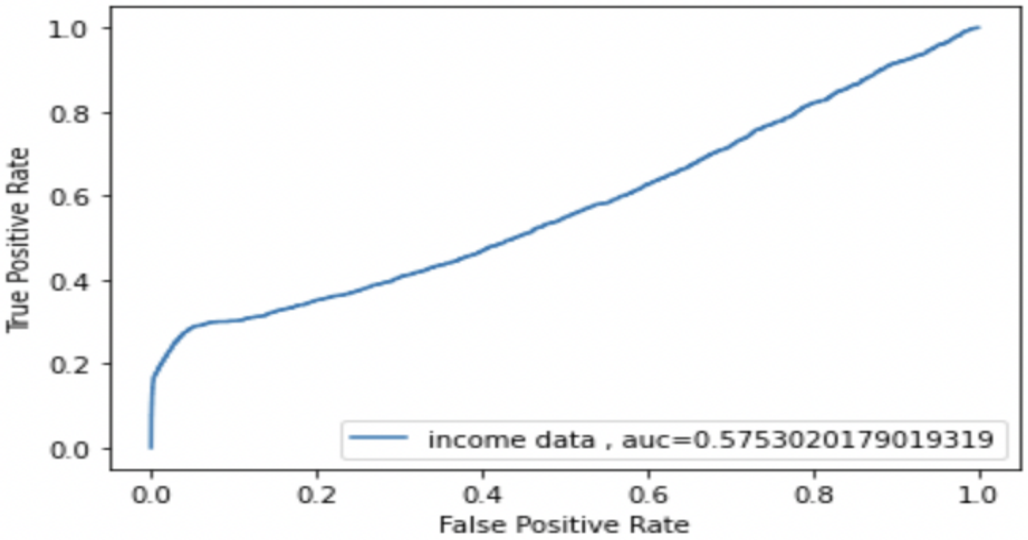


Figure 5 (second model):



As one can observe from the results, the AUC value for the first model was about 0.71 while the value for the second model was about 0.58. As the AUC from the first model returns a much higher value than the second model, it implies that it is the better model for classification. This was quite surprising, given the results of the confusion which seem to imply that the second model had better prediction than the first model.

Conclusion:

The following paper examined the education level variable to assess its effectiveness in classifying individuals based on their income level based on the Census Income Data from the UCI Machine Learning Repository. Two separate logistic regression models were created for classification, one which only uses the education level variable, while the other utilizes several other variables. Confusion matrices were made for both models from which ROC curves were constructed. The AUC values for both models were compared to reveal that the first model, which only contains the education level as the explanatory variable, performed better than the second model. This could be explained by noise caused by variables in the second variable which don't have much relevance in the case of study. An example of such a variable could be "capital gain" and "capital loss" as investment results can be seen as independent of the level of income of an individual. While it was concluded that the first model is the better model, both models do not have very high values, indicating that neither would make great classification models. However, model 1 does show that the education level does play a significant role in determination of income level. For future research, feature engineering could be further explored to build the best possible model for classification of income level using a combination of variables. In addition, the causal inference can be studied to examine whether the education level and income classification have a causal relationship.

Link to google collab:
https://colab.research.google.com/drive/1tqiZpPuf93kbMHcE56i7gJr3ApvG34Kg?usp=sharing

List of References:

Navlani, A. (2019, December 16). *Python logistic regression tutorial with Sklearn & Scikit*. DataCamp. Retrieved November 20, 2022, from https://www.datacamp.com/tutorial/understanding-logistic-regression-python

Ucfaibot. (2020, September 7). *Core-FA19-regression*. Kaggle. Retrieved November 20, 2022, from https://www.kaggle.com/code/ucfaibot/core-fa19-regression/data

Wolla, S. A., & Sullivan, J. (n.d.). *Education, income, and wealth*. Economic Research - Federal Reserve Bank of St. Louis. Retrieved November 20, 2022, from https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth#:~:text=Education%20is%20often%20referred%20to,incomes%20(see%20the%20table)