4.

For the prediction portion of this assignment, we used the following methods/objects from the sklearn library: LinearRegression, train\_test\_split, classification\_report, confusion\_matrix, and accuracy\_score.

Our exploration of the data using the various visualizations did not reveal any obvious correlations between the fields, so we first tested training a linear model using two fields that intuitively would make sense as relevant: age and education. Then, to see if we could make it better, we adding more and more fields that once again, intuitively made sense. Below are the results for each, we used used sklearn’s train\_test\_split to generate training and testing sets at the various breaks requested (50/50,60/40,70/30). In each table, marked by the fields used in the model, you will see the accuracy and weighted f1-score for the 3 splits requested. The results seem to demonstrate that the model performs best with all fields included, getting a weighted f1-score of 0.8. This implies that the model’s performance depended not so much on which fields were used, but the number of fields used, which would make sense with our visualizations showing no clear correlations; the model does not have clear linear relations so it performs best when the most information is provided. Below are these initial results, after which we will show further exploration using the visualizations, which shows that it’s not just about the number of fields; some fields are actually more relevant than others and are responsible for most of the results.

|  |  |  |
| --- | --- | --- |
| **['age','education']** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.76** | **0.67** |
| **60/40** | **0.76** | **0.67** |
| **70/30** | **0.77** | **0.67** |

|  |  |  |
| --- | --- | --- |
| **['age','education','race']** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.76** | **0.67** |
| **60/40** | **0.76** | **0.67** |
| **70/30** | **0.76** | **0.67** |

|  |  |  |
| --- | --- | --- |
| **['age','education','race','sex', 'hours-per-week']** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.77** | **0.7** |
| **60/40** | **0.77** | **0.7** |
| **70/30** | **0.77** | **0.7** |

|  |  |  |
| --- | --- | --- |
| **[all fields]** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.83** | **0.8** |
| **60/40** | **0.82** | **0.8** |
| **70/30** | **0.83** | **0.8** |

At this point we decided to use the visualizations as much as we could. Although there did not appear to be any clear linear relationships, some field pairings did demonstrate a loose trend. To begin, we tested fnlwgt with capital-loss and did obtain better results than our initial two-field test. Then, by adding another promising field, in this case capital-gain, the results jumped to almost what they are when all fields are included. This clearly shows that there is a subset of fields more responsible for the results, and the all fields results are most likely just a reflection of that subset. For instance, adding age did essentially nothing to these results, so it is likely doing nothing in the all fields test either. Further studies of this dataset could use more in-depth data analysis to find the exact subset responsible for the highest results; this study is confident the fields fnlwgt, capital-loss, and capital-gain would be included in that subset.

|  |  |  |
| --- | --- | --- |
| **['fnlwgt','capital-loss']** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.78** | **0.71** |
| **60/40** | **0.78** | **0.71** |
| **70/30** | **0.78** | **0.71** |

|  |  |  |
| --- | --- | --- |
| **['fnlwgt','capital-loss','capital-gain']** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.81** | **0.78** |
| **60/40** | **0.81** | **0.78** |
| **70/30** | **0.81** | **0.78** |

|  |  |  |
| --- | --- | --- |
| **['age','fnlwgt','capital-loss','capital-gain']** |  |  |
|  |  |  |
|  | **accuracy** | **f1-score** |
| **50/50** | **0.81** | **0.78** |
| **60/40** | **0.81** | **0.78** |
| **70/30** | **0.81** | **0.78** |

Included here is also the full results for the fnlwgt, capital-loss, capital-gain test. It is worth noting that for all tests the model performs much better on predicting someone is below the 50k threshold (class 0) than above (class 1). This is because the data has much more class 0 data to train on. Future studies could use a method such as SMOTE to balance the data more and provide a better training ground for class 1 data points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **['fnlwgt','capital-loss','capital-gain']** |  |  |  |  |
|  |  |  |  |  |
|  | **precision** | **recall** | **f1** | **support** |
| **0** | **0.82** | **0.96** | **0.89** | **7550** |
| **1** | **0.71** | **0.3** | **0.42** | **2219** |
| **accuracy** |  |  | **0.81** | **9769** |
| **macro avg** | **0.77** | **0.63** | **0.66** | **9769** |
| **weighted avg** | **0.8** | **0.81** | **0.78** | **9769** |

The results for class 0 predictions are actually very, very good. Improving the class 1 results would improve the overall performance of the model significantly.