Adult Dataset:

|  |  |  |
| --- | --- | --- |
| Features | Possible Values | Usefulness |
| Age | Age values given in years; range is whole numbers greater than 0 | Yes, income is dependent on experience in the workforce which correlates with age |
| Workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked | Yes, income is also dependent on type of job. |
| Fnlwgt | Whole numbers greater than 0 | No, the amount of the population represented by a person doesn’t seem relevant while classifying their income range. |
| Education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool | No, redundant data to education-num |
| Education-num | Whole numbers greater than 0 and less than 16 | Yes, education is highly correlated with income. |
| Marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse | Yes, home-life and relationships are important in people’s lives which also might influence their income |
| Occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces | Yes, Occupation is mostly directly correlated with income. |
| Relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried | Maybe, seems similar to marital status but could have an impact on classification |
| Race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black | Maybe, income varies greatly with race but since there is such a large number of white people, it end up hurting our model |
| Sex | Male, Female | Yes, income also varies with sex |
| Capital Gain | Positive monetary values | Yes, wealthy people are more likely to invest |
| Capital Loss | Positive monetary values | Yes, same logic as Capital Gain |
| Hours/week | Positive whole number values | Yes, income and labour hours are correlated |
| Native Country | Names of countries | Yes, immigrants from certain countries are more likely to be in similar lines of work |

We decided to go with the Adult dataset for our classification set. The reason we chose this set was because it had only two classes to distinguish between, and most of the features seemed relevant in sorting samples. The zoo dataset had primarily Boolean features that corresponded to whether the animal exhibited a certain feature which would have made classification a lot easier. The reason we didn’t pick the zoo set is because we didn’t feel like there were enough samples in the set to create an accurate model. Considering there are millions of animal species in the world, with each having unique sets of features and characteristics, 79 samples seemed too few.

One drawback of using the adult dataset is that most of the features are nominal, which means that they will need to be converted to numerical data. The features are based on general information about a person including questions about their demographic information, education, occupation, home-life, and investment gain/loss. We decided that fnlwgt and education would not be useful in our classification. Fnlwgt is the estimated amount of the population that a person in the data represents. We didn’t think this to be relevant because it wasn’t a direct attribute of the person itself. The reason we decided to not use education is because the education number was already a numerical counterpart to education with different education levels corresponding to different numbers. There were a few samples with missing features, which we decided to remove as there were an insignificant number of them.

Wine Dataset:

|  |  |  |
| --- | --- | --- |
| Features | Possible Values | Usefulness ( all of them had similar explanations, see in paragraph below) |
| Fixed acidity | Positive numbers with one decimal place | Yes |
| Volatile acidity | Positive numbers with two decimal places between 0 and 1 | Yes |
| Citric acid | Positive numbers with two decimal places between 0 and 1 | Yes |
| Residual Sugar | Positive numbers with one decimal place | Yes |
| Chlorides | Positive number with three decimals around .100 | Yes |
| Free Sulfur Dioxide | Positive whole numbers in the tens | Yes |
| Total Sulfur Dioxide | Positive whole number ranging from 18-300 | Yes |
| Density | Positive numbers with up to four decimal places close to 1.0000 | No, there seems to be very little variance with density, and the variance looks to be random |
| pH | Positive numbers with up to two decimal places close to 3.00 | Maybe, seems important but all values are close to 3.00 |
| Sulphates | Positive numbers with up to two decimal places from 0-1.00 | Yes |
| Alcohol | Positive numbers with up to one decimal place in the range of 8-14 | Yes |

For the regression dataset we decided to go with the wine dataset. The reason we picked this dataset is because it had all numerical features and most of the features seemed important. We decided against the student dataset because it had significantly more features, with a lot of more nominal features. It would have taken a lot longer to sort through all the features and determine which ones are important vs which ones are not. One thing we will have to change about the wine dataset is the scale of its features are wildly different, so we will have to convert them to a standardized scale. We also noticed that a lot of the wines in the test set were clumped towards the middle for quality. Due to this, it might be more difficult to classify average wines compared to high- and low-quality ones.

We determined that all the features except density were important in regression. We deemed density not important because all the values were very close to 1.000, which they should be because primarily water-based liquids will all have similar densities. The other features all seemed like they would be important because they had more variance and dealt with chemicals which could directly affect the taste of the wine.

Results:

For the adult dataset, we converted all nominal features into booleans using pd.get\_dummies() so we had numerical data to work with. For missing features, we assigned the most common ones to the ones that were missing in the test and training sets even though we had initially planned to omit samples with missing features completely. For example, we set people with missing native countries as United States, and people with missing work class as private. We used Naive Bayes and kNN classifiers with odd number neighbors from 1 to 31. At the beginning we decided to scale our originally numerical features using pre-processing and the min\_max scaler so that all our feature values were scaled from 0 to 1. We omitted the fnlwgt and education as we didn’t think they were very helpful in classification. When testing our classifiers, we used cross validation with five folds. Initially we found accuracy scores for each classifier and picked one of the best(n=23), but it ended up not performing that well on the test set. After not doing well on the test set, we sought to figure out why our accuracy was as low as it was.

Our first thought was that we had too many features and our model was overfitting. We decided to experiment with different feature options first. In our second submission, we decided to omit the relationships feature because we began to question whether or not it had relevance to the classification. We used the same kNN model with 23 neighbors and found that it did slightly better than before, but still pretty poorly overall. We then decided to reevaluate why we scaled our numerical features to between 0 and 1 in the first place. After un-scaling our numerical data, we tested our classifiers again using cross validation with five folds and received significantly better results than we did before. We decided to use the naive bayes model for our third submission which did significantly better, but still not as good as we’d liked. For our final model, we decided to omit race because of the skew towards white people in the data. We found that kNN with 19 neighbors gave us the best accuracy scores. After applying it to the test set and uploading our submission, we were happy with the accuracy we got in the end. At this point, we tried to tinker with our features further but we were unable to improve the accuracy scores from kNN with 19 neighbors.

For the wine dataset, we used pre-processing and the min\_max scaler to scale all our feature values from 0 to 1. We did this because some features were in the thousandths while others were in the tens and hundreds. We wanted to make sure that each feature would hold similar weight during regression. After converting our train and test features to the new scale, we tested a variety of models to see which one would be the best. We tested linear regression as well as kNN regression with odd number values of k from 1 to 15. We used cross-validation using the training set with 10 folds to obtain the RMSE for each model which we compared to determine the best model. While kNN with 15 neighbors showed the lowest RMSE value at .77, we decided to go with linear regression which had an RMSE value of .78. The reason we picked linear regression over kNN is because we feared that the high value of k would lead to a model that was too complex. We uploaded the predictions from linear regression, kNN n=15, and kNN n=9 to Kaggle and saw that linear regression gave us the best score. At this point, we experimented with our features and retrained our models with a different blend of features and found that it didn’t improve the RMSE values that well. In the end we stuck with linear regression with all the features except density.