Adults_Dataset_Analysis copy

June 7, 2022

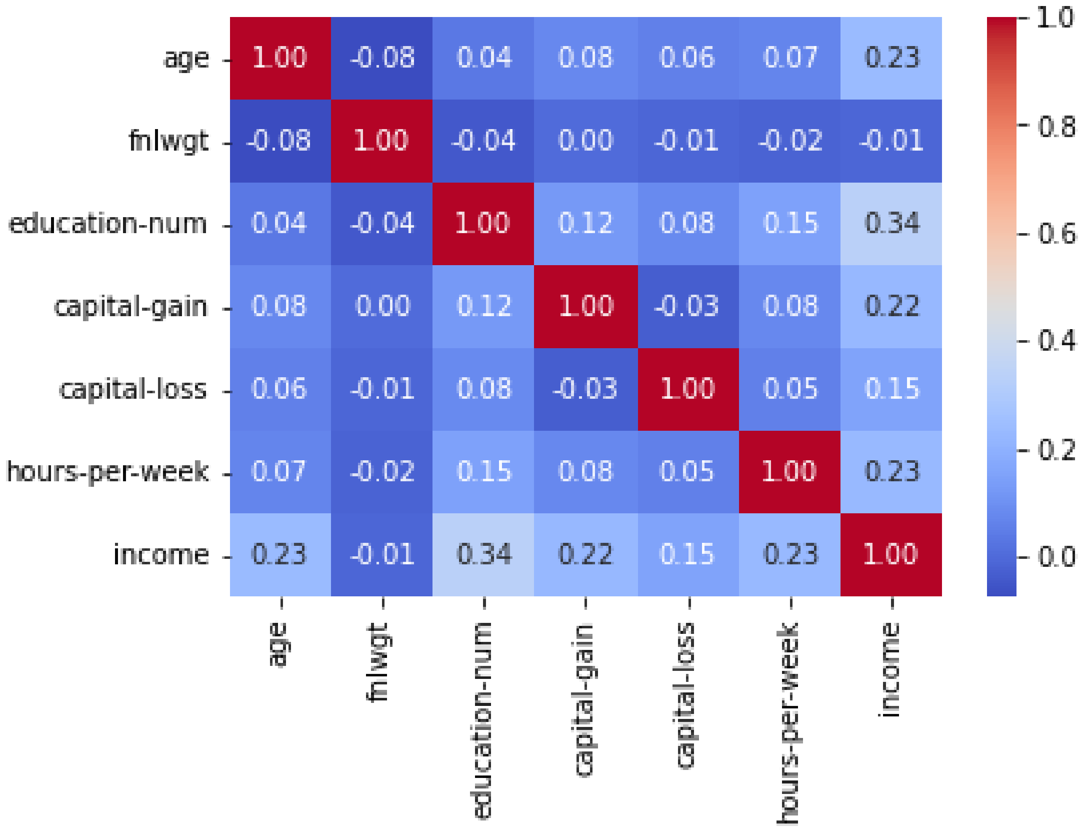
1 Classification with the Adult dataset

1.0.1 Author: Marcel Rodrigues de Barros

1.0.2 NUSP: 5947197

[64]:

```
g = sns.heatmap(adults[numeric_features].corr(),annot=True, fmt = ".2f", cmap =␣
 ↪"coolwarm")
```



1.0.3 EDA

Age, education, capital-gain, capital-loss and hours-per-week are all positively correlated with the income label.

Kernel density plots shows that it is much more common (relative to sample size) for white men to earn more than 50K than non-white and women.

1.0.4 Feature engineering

Removals

- 'fnlwgt' field is a weight that represents how many people can be described by the respective row. It does not correlate to 'income' as shown above and can be discarded.
- 'education' is just the text version of 'education-num' and can also be dropped.

Auxiliar columns

- 'capital-loss' and 'capital-gain' are mostly zeros.

Encoding and Normalization Categorical feature were encoded using one-hot encoding and numeric features were normalized using Z-score normalization.

1.0.5 Models and Hyperparameters

All hyperparameters were defined using KFold cross-validation with k=5. Please check the full report for details.

Classifier	Parameter	Value
KNN	n_neighbors	26
Random Forest	n_estimators	190
Gaussian NaiveBayes	var_smoothing	0.1
Logistic Regression	penalty	l2
MLP	layer configuration	[40,40]

1.0.6 Results

Classifier	Accuracy	Precision [<=50K,>50K]	Recall[<=50K,>50K]
KNN	0.852	[0.89,0.72]	[0.93,0.61]
Random Forest	0.850	[0.91,0.58]	[0.83,0.74]
Gaussian NaiveBayes	0.811	[0.88,0.74]	[0.93,0.60]
Logistic Regression	0.855	[0.91,0.58]	[0.93,0.61]
MLP	0.860	[0.88,0.76]	[0.94,0.59]

1.0.7 Conclusion

As shown, the classification using one-hot encoding for categorical features and Z-score scaling for numeric feature achieve above 85% accuracy for all classifiers, except for NaiveBayes.

Feature analysis show the discrepancy between income for women vs men and for white vs non-white citizens.