HW5 – Option 'A'

The question is asking to predict a person's "empathy" on a scale from 1 to 5 given the dataset.

Data Preprocessing

- Convert all categorial data to numerical values.
- Feature reduction by converting weight and height to BMI and dropping height and weight columns.
- Performed binning to convert continuous data like age, siblings and BMI into categorical data.
- Can't drop all records with missing values as in total 336 rows with missing values are present so dropped rows for which empathy value is missing as we cannot manipulate label.
- For the remaining missing data, took the mean of the column and inserted it at empty places where mean will denote the average of the column.
- Since the data size is less, I divided the data as 85% training data and rest as testing data.
- I took out the empathy column from the dataset which will be used as the label afterwards.
- Since there are 150 features and size of data is 1010, the features to records ratio is high, so we need to perform feature reduction.
- To reduce features, I used sklearn's feature selection class SelectKBest to select the top 40 features according to the best 40 scores.

Model Building

- Now, after preprocessing I ran base qualifier "Most frequent class". The most frequent rating that I got was "5" for y-label. This was giving an accuracy of 36.72%. So, I must make a classifier to give better accuracy than this.
- I have used 9-fold cross validation for all classifiers to get better results, as the dataset is small, and test data and training data may have high variance, which can give inaccurate results, also I have got results without cross-validation by training on 85% data.
- The first classifier I tried was K-Nearest neighbor and for this k was set by default to 5 and we have 5 ratings, so I didn't change K. I thought it will be good to use KNN since we have multiclass and so dividing them into 5 clusters will do the work. But, it gave a cross-validation accuracy of 31.84% which is less than the base classifier accuracy, so I left it.
- Then I tried Decision Tree and it gave me an accuracy of 37.12% which was still less.
- Multinomial and Gaussian Naïve Bayes gave 39.03% and 40.62% accuracy respectively.
- I used logistic regression with multinomial parameter as we have multiclass, and it gave me 42.00% accuracy and Random Forest gave 42.72% with depth 6.
- I felt XGBoost will give best result since it is an ensemble, so I installed XGBoost and ran it and as expected XGBoost gave me the best result until now, i.e. 45.30%.
- Finally, I used RBF kernelized SVM as we have 5 classes. To tune it, I used GridSearchCV and passed Cs and gammas to it to select the best C and gamma. Then I passed the C and gamma returned from it to kernelized svm which gave me 45.79% accuracy which showed better results than base classifiers.

BitBucket URL:
https://bitbucket.org/piyush910/imlyoungpeoplesurvey/src/master/