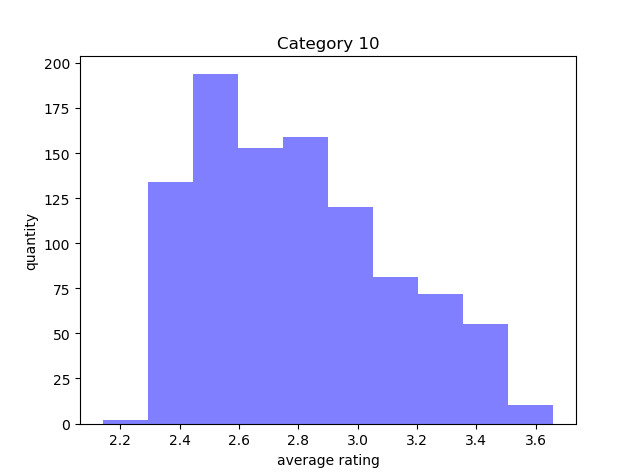
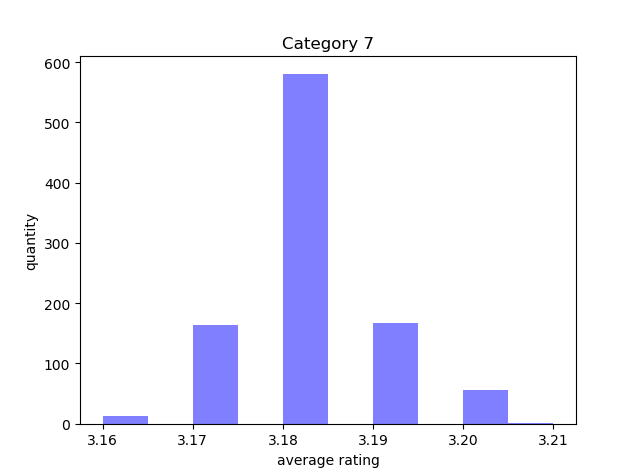
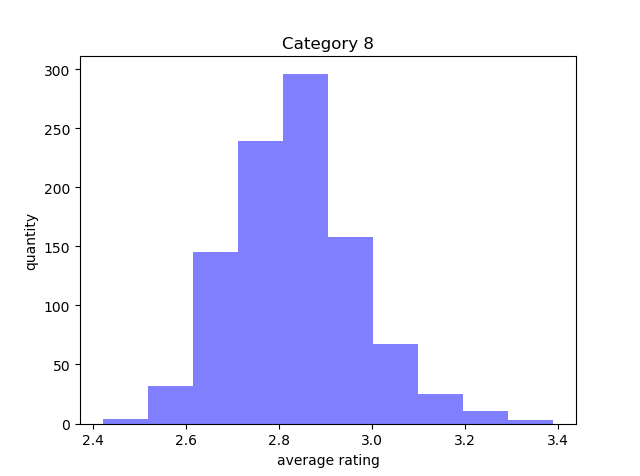
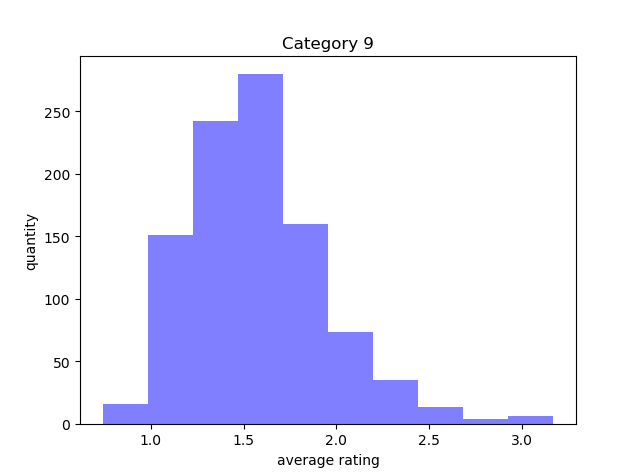
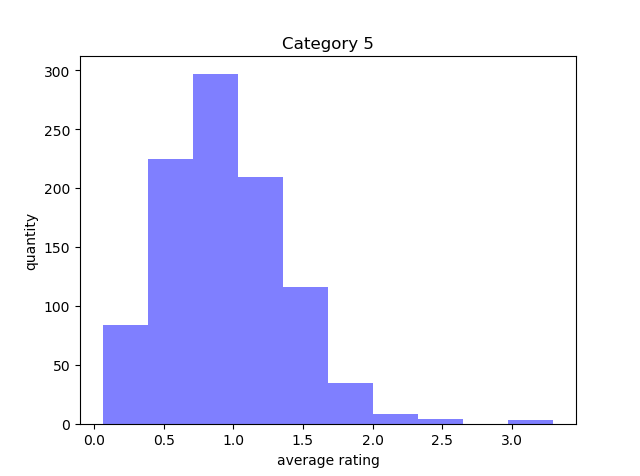
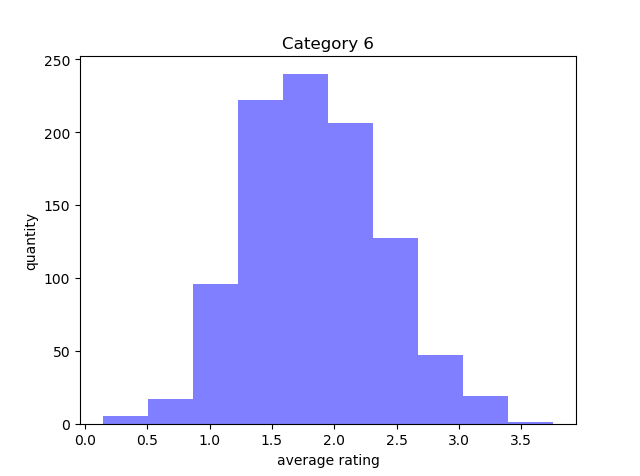
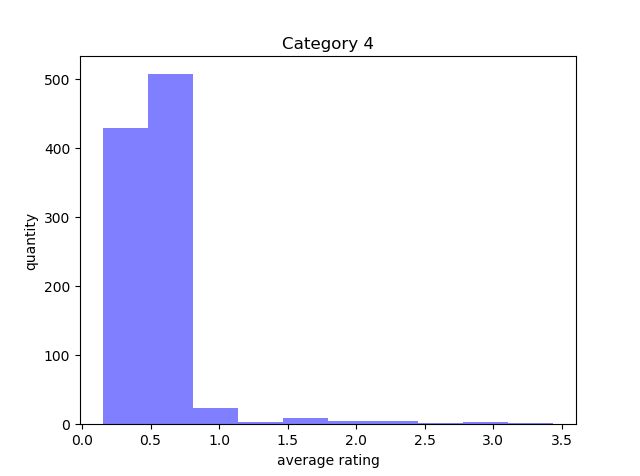
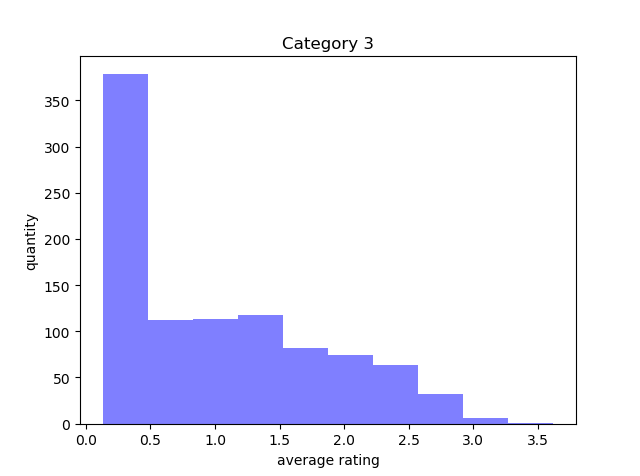
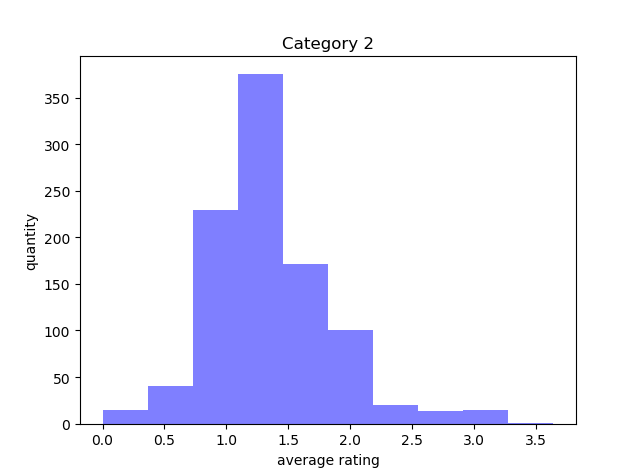
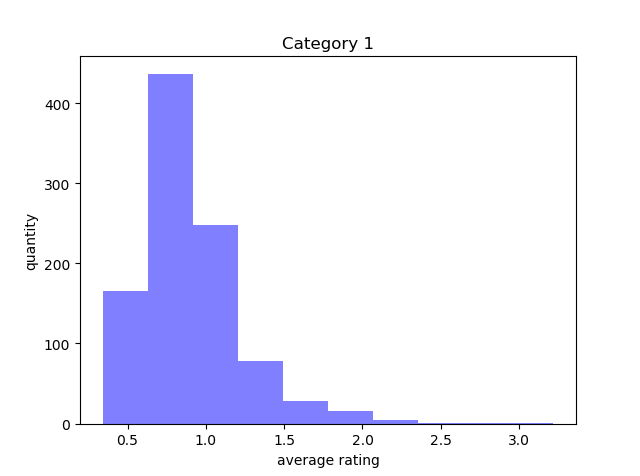
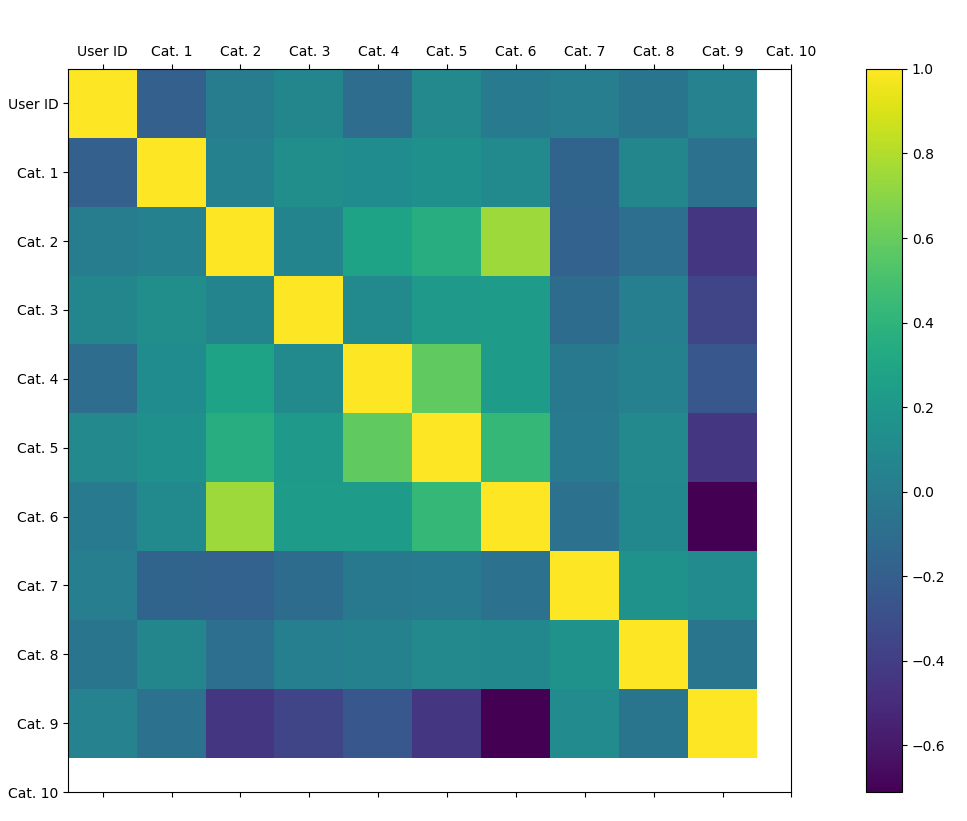
UFPR - Universidade Federal do Paraná  
SEPT - Setor de Educação Profissional e Tecnológica  
Especialização em Inteligência Artificial Aplicada  
IAA006 - Arquitetura de Dados  
Prof - Dieval Guizelini  
Nome - Hélinton P. Steffens

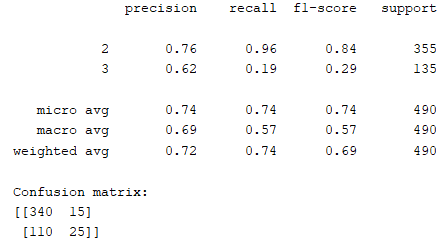
Base de dados: [Travel Reviews](https://archive.ics.uci.edu/ml/datasets/Travel+Reviews)  
Número de atributos: 11   
Número de classes: 4 (Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0))  
Número de instâncias: 980  
Objetivo: Classificar as pessoas que gostaram ou não do passeio de categoria 10 com base nas avaliações dos outros passeios.  
Os dados estão distribuídos conforme apresentado nos histogramas abaixo:



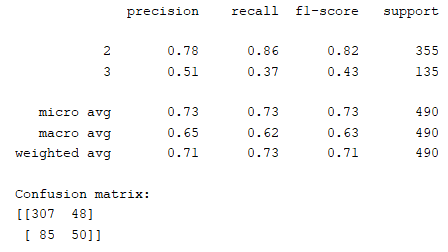
Sendo que a correlação entre as colunas pode ser observada pelo mapa de calor das correlações:

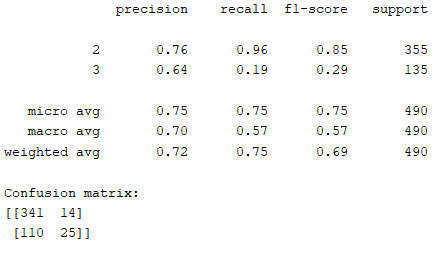


Outro ponto a ser atentado é que nesta base de dados, a coluna User ID é um campo sequencial e que a base não possui valores inválidos, ou seja, não possui nulos ou N/A e nem valores fora dos limites validos de 0 a 4. Ao ser executado os algoritmos random forest, SVM e perceptron foi obtido os seguintes resultados sem ser realizado pré-processamento da base, sendo que a única técnica aplicada foi bootstrap para dividir a base de dados:

 SVM

RandomForest





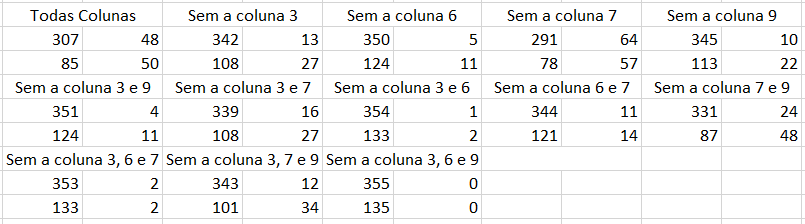
MLP

Ao ser reanalisado as correlações em relação à informação a ser descoberta temos as seguintes situações, primeiro podemos observar as seguintes correlações:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cat. 1 | Cat. 2 | Cat. 3 | Cat. 4 | Cat. 5 | Cat. 6 | Cat. 7 | Cat. 8 | Cat. 9 | Cat. 10 |
| 0.050 | -0.065 | -0.440 | -0.352 | -0.247 | -0.438 | -0.710 | 0.114 | -0.045 | 1.000 |

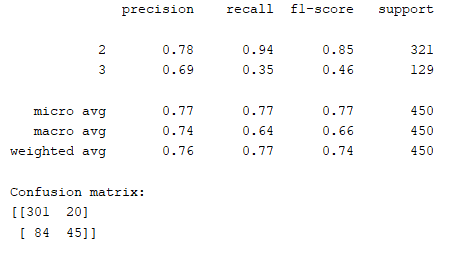
Ao tentarmos remover as colunas com a correlação negativa obtivemos uma melhora na classificação dos dados conforme pode ser observado nas matrizes de confusão abaixo:

Matrizes de confusão – Sem as colunas com correlação negativa



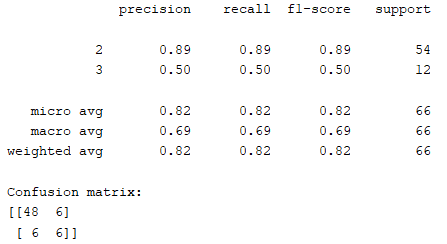
Como pode ser observado ao ser removido as colunas 7 e 9 houve uma redução nos dados classificados de forma errônea de 133 para 111, ao remover as colunas 3,7 e 9 também obtivemos uma melhora considerável caindo os dados classificados de forma errada para 113. Será optado por seguirmos sem as colunas 7 e 9, que apresentaram uma classificação mais assertiva, apesar de que sem as colunas 3,7 e 9 teríamos um desempenho maior por estarmos trabalhando com menos dados e a perda de informação ser pequena neste caso, mas como o dataset é pequeno será optado pelo resultado mais assertivo. Ao remover a coluna User ID que é sequencial e os valores duplicados da base conseguimos uma melhora na classificação conforme pode ser observado na imagem abaixo.

SVM – Sem valores repetidos e as colunas 7 e 9



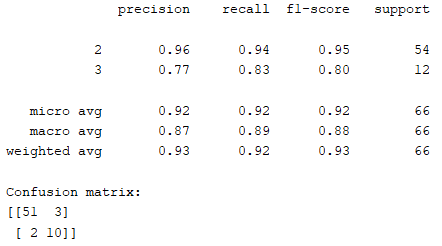
Para tentar minimizar os efeitos do desbalanceamento das classes, foi optado por utilizar a técnica Repeated K-Fold, com os seguintes parâmetros: n\_splits=15, n\_repeats=30, random\_state=12883823. Com isto foi obtido uma melhora significativa conforme pode ser observado na imagem abaixo.

SVM – Repeated K-Fold

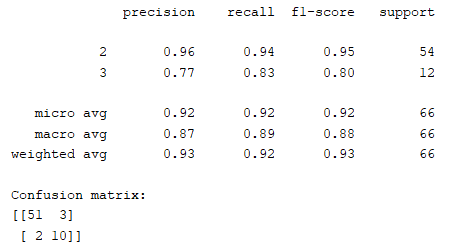


A próxima estratégia utilizada foi tentar padronizar os dados utilizados, utilizando StandardScaler, disponibilizado pelo Python, e com isto temos valores normalizados, ou seja, tendo a subtração da média de cada característica e então a dividindo pelo desvio padrão. Com isto foi possível obter uma melhora no modelo, como pode ser observado na imagem abaixo.

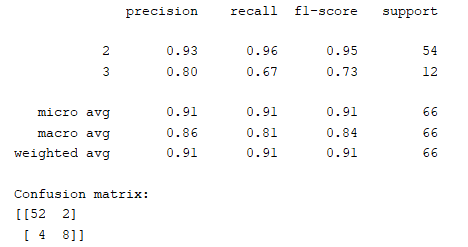
SVM – Com os dados padronizados



Para definir que um modelo é melhor do que o outro, foi utilizado a métrica micro avg. Nas imagens abaixo será mostrado os resultados dos algoritmos randomForest e MLP, com os mesmos tratamentos realizados anteriormente.

RandomForest

MLP



Por fim, é possível concluir que os resultados dos três modelos finais melhoram, em relação aos modelos inicias, que não possuem nenhum pré-processamento. Então levando em conta a métrica utilizada para avaliar a qualidade de um modelo, que neste caso foi a micro avg, foi alcançado um total de 92% de precisão em relação aos 75% obtidos inicialmente no melhor modelo que tinha sido a MLP. Os resultados finais dos três algoritmos foram bem semelhantes, tendo a MLP e SVM tido a mesma taxa de precisão, enquanto o randomForest foi levemente inferior com 91% de acurácia.