Report: Insights Gained from SHAP Analysis on Chocolate and Mushroom Datasets

The Random Forest and Decision Tree models' performance on two distinct datasets was clarified through the use of SHAP (SHapley Additive exPlanations), which offered important insights into feature significance, model correctness, and potential enhancements. We supported the objectives of Explainable AI (XAI) by employing SHAP to closely analyze model behavior and improve the transparency of their decision-making processes.

1. Key Feature Importance and Model Decision Pathways

The most important features in each dataset's classification tasks were shown by SHAP summary graphs. In order to classify data points in the chocolate dataset as "user" or "non-user," attributes with higher SHAP values were essential. This allowed us to determine the main factors influencing each model's predictions. In the mushroom dataset, the Decision Tree model made judgments using simpler, more direct pathways, whereas the Random Forest model depended on a number of influential characteristics and intricate interactions. This investigation demonstrated how well the Random Forest handled complex data patterns, something the Decision Tree model was less able to achieve.

2. Examining Model Prediction Differences

SHAP force plots provide a more thorough examination of the logic underlying each forecast, particularly the inaccurate ones. These plots showed how particular feature values could affect predictions toward or away from the right classification for both datasets. The force plots enabled us to identify the aspects that might have misled the models when they made mistakes, potentially as a result of biased or noisy data. This was especially noticeable in the Decision Tree model, which appeared to be more prone to errors due to its excessive dependence on individual variables, indicating regions that needed to be adjusted.

3. Identifying Conditions for Success and Failure

Plots of dependencies and interactions showed the strengths and weaknesses of each model. Both models demonstrated reduced accuracy in the chocolate dataset when faced with extreme or sparse feature values; however, the Random Forest model fared better in these difficult situations. The mushroom dataset showed a similar trend, with the Decision Tree model performing poorly in areas with intricate feature interactions. While the Decision Tree struggled to manage more complex dependencies, this comparison study demonstrated how resilient the Random Forest was in managing a variety of feature relationships.

4. Assessing Model Trust and Reliability

SHAP analysis also allowed us to evaluate each model’s trustworthiness. Predictions that were consistent with expected feature importance increased confidence in the model’s reliability. In both datasets, the Random Forest model’s balanced reliance on relevant features enhanced its credibility, while the Decision Tree sometimes focused on less meaningful features, which suggested potential overfitting. These findings emphasized the importance of model selection based on task complexity and the need to ensure that a model's feature reliance aligns logically with the data.

5. Potential Enhancements for Model Performance

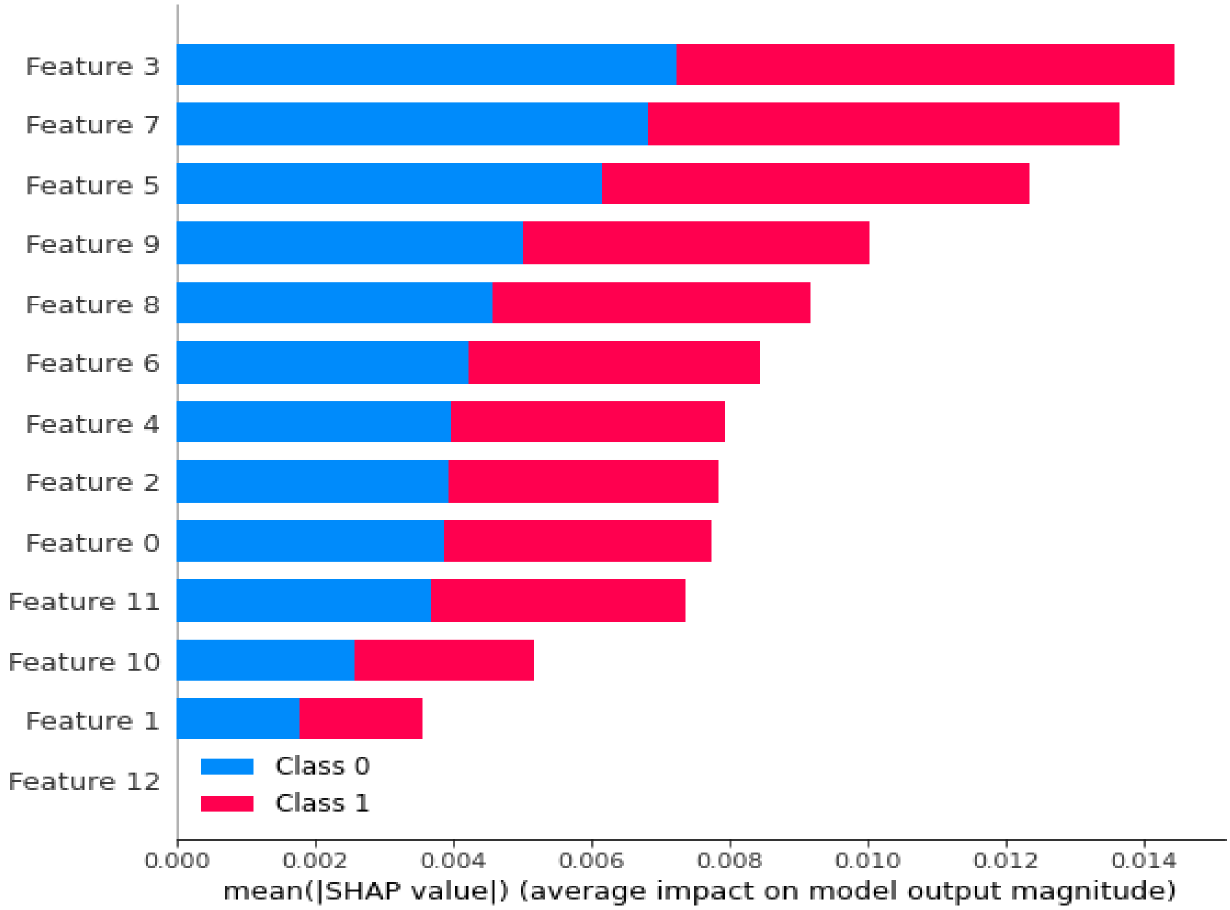
The SHAP findings suggested several improvements. For example, in both datasets, feature engineering could be refined to reduce the impact of highly variable features that disproportionately influenced specific subgroups. Additionally, SHAP revealed that using more advanced models, such as gradient boosting, might better capture subtle feature relationships, which could improve both accuracy and generalizability in future analyses.

Conclusion

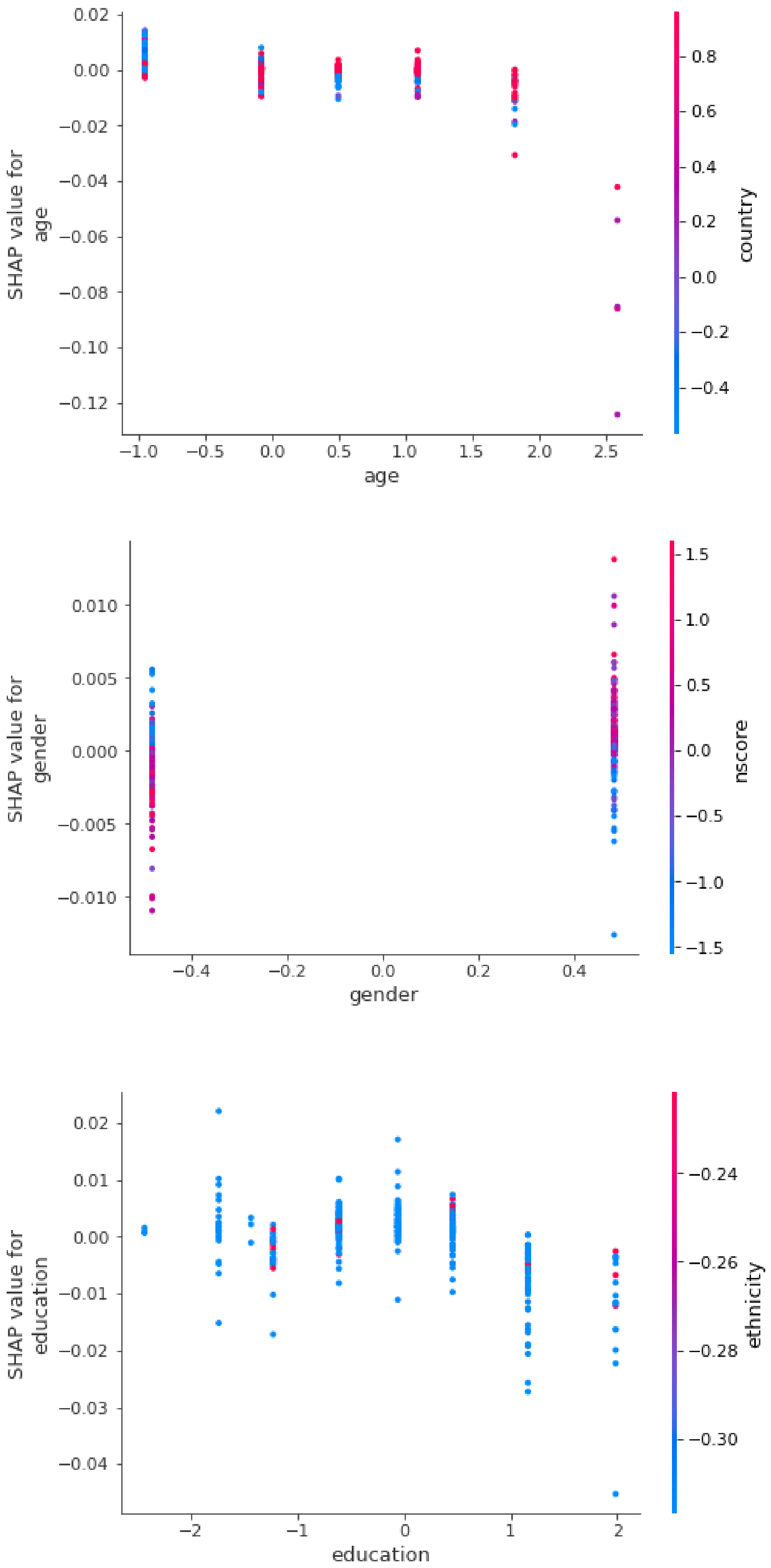
Overall, the SHAP analysis offered insightful information on how the Random Forest and Decision Tree models function. SHAP proven to be a crucial tool for creating dependable, interpretable models by exposing the impact of various features and pointing out regions where models worked or didn't. This method helps guarantee that models not only function accurately but also fit in with clear and intelligible decision-making procedures.

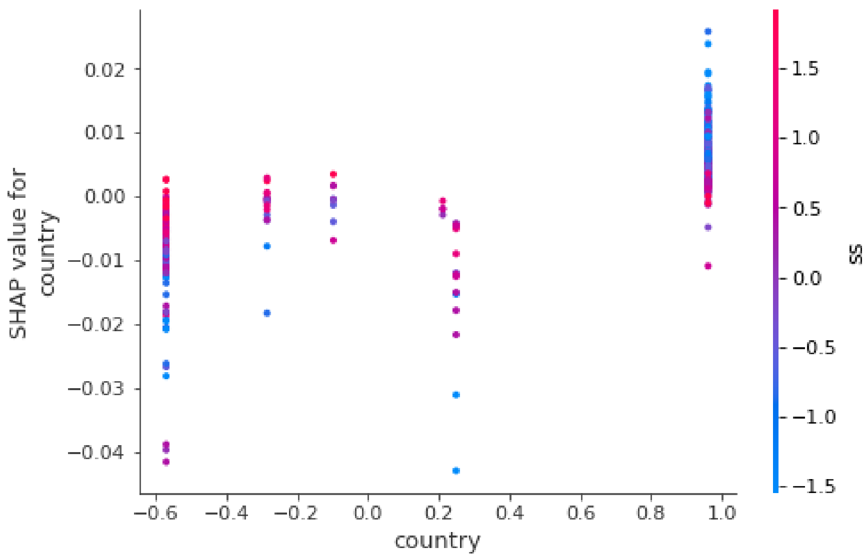
PLOTS – Computing SHAP values with the best model (Random Forest) on Chocolate test set

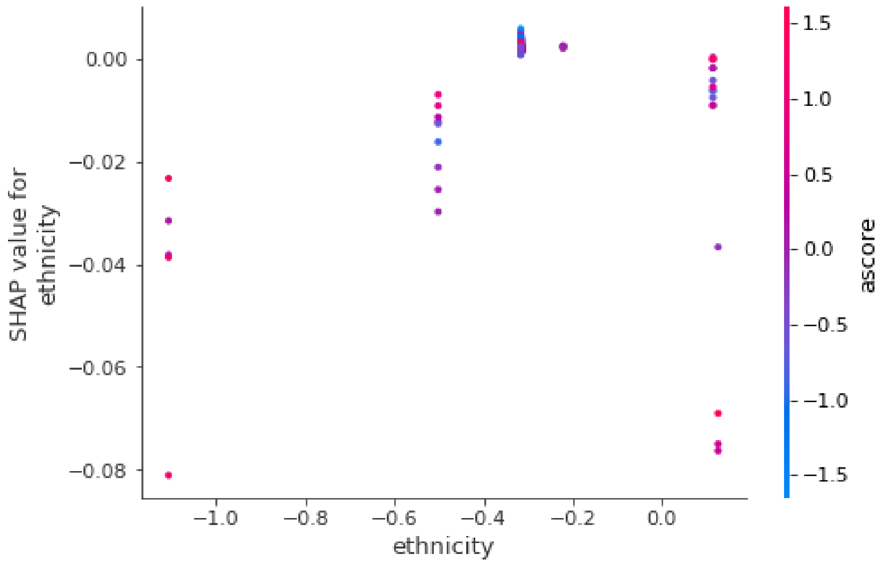
```
SHAP Summary Plot for Model Random Forest on Dataset Chocolate Test Set
```

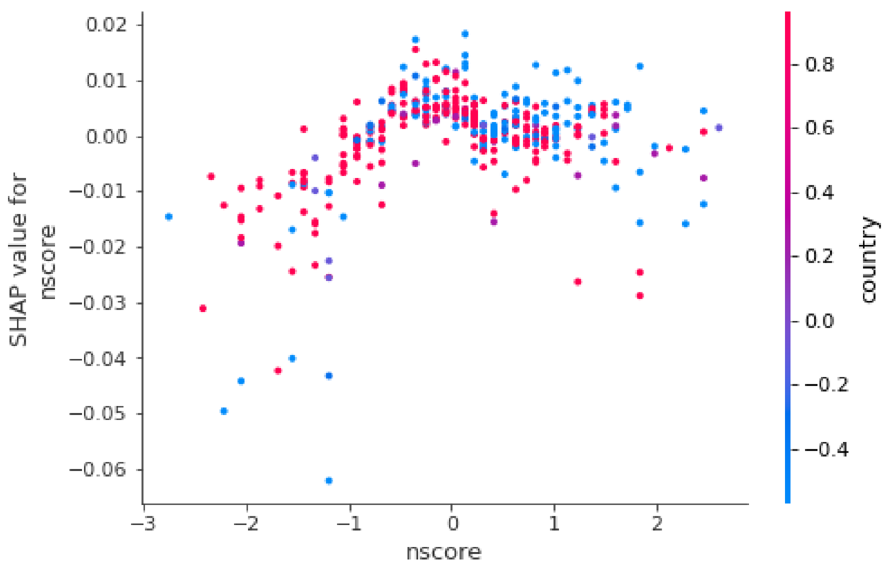


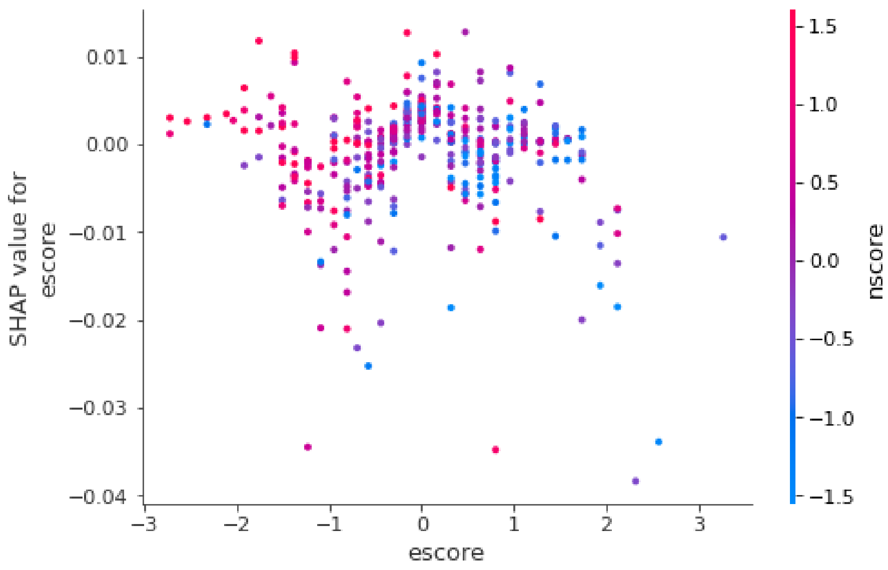
SHAP Dependence Plot on Model Random Forest with Dataset Chocolate-TEST

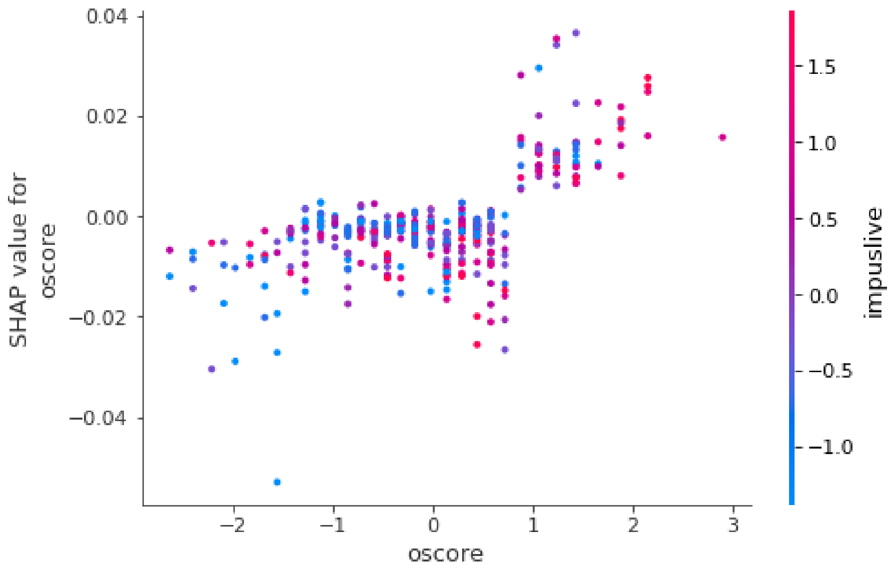


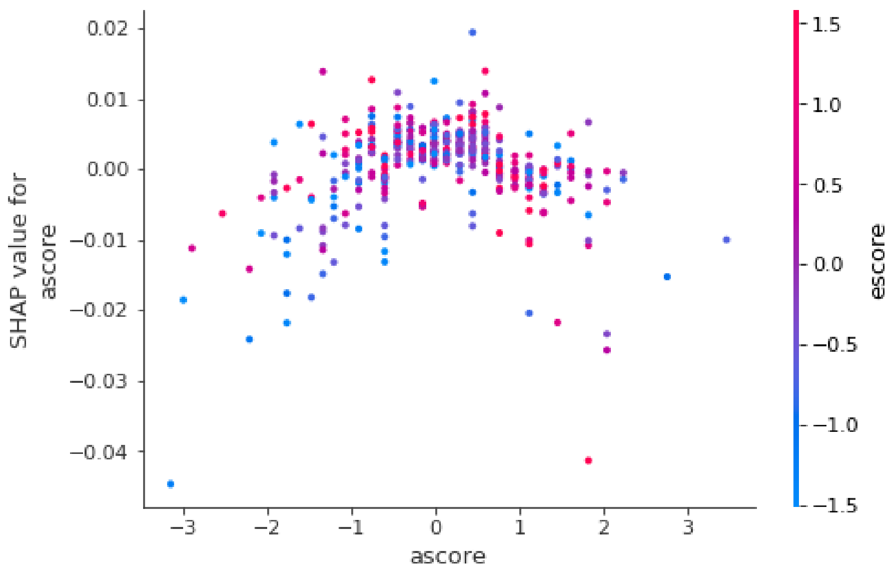

SHAP value for country
country
ss

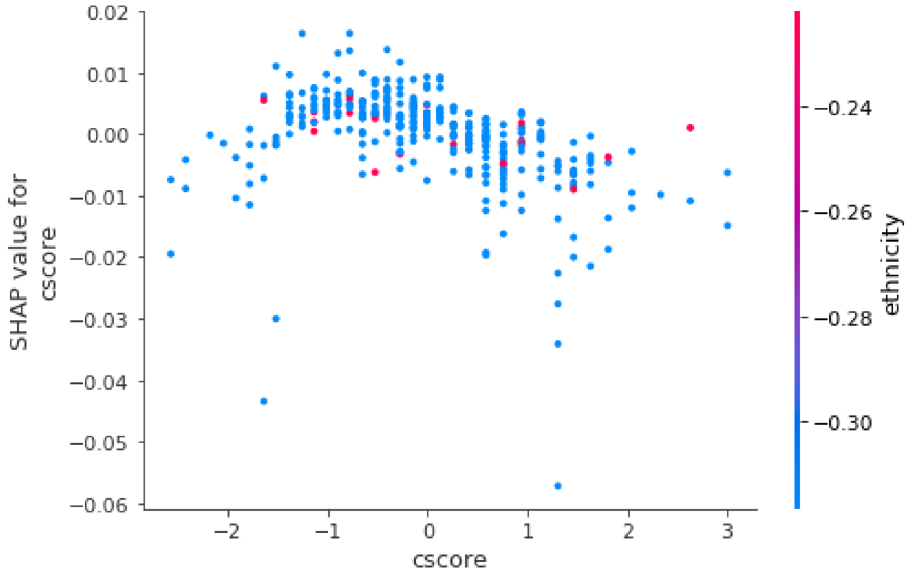

SHAP value for ethnicity
ethnicity
ascore

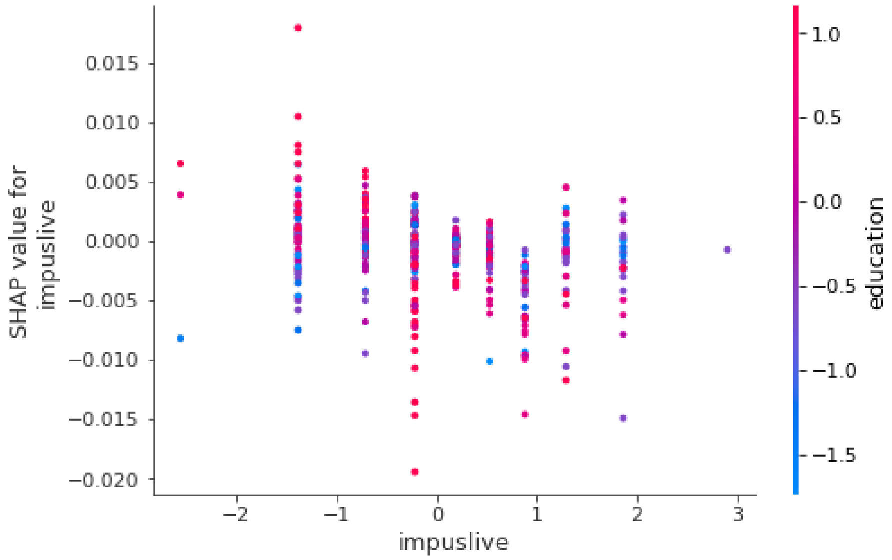

SHAP value for nscore
nscore
country

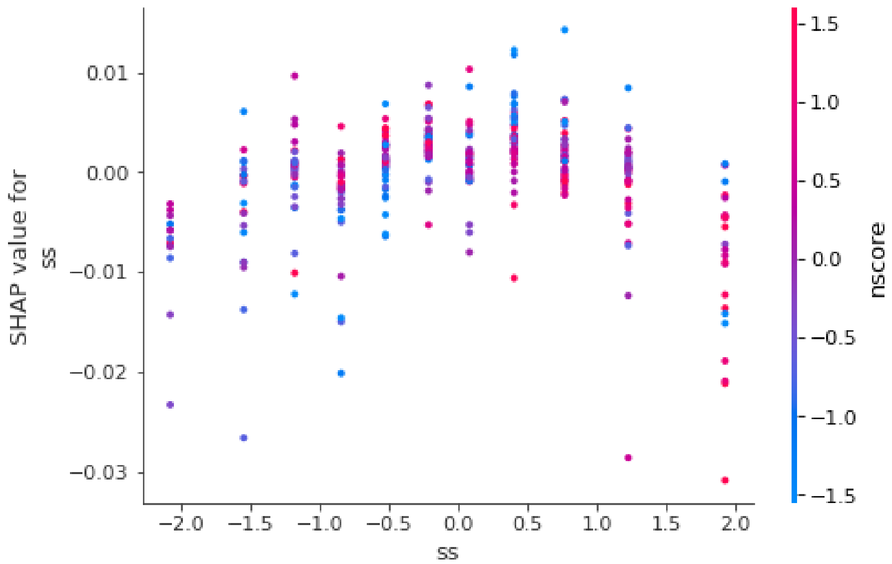

SHAP value for escore
escore
nscore


SHAP value for oscore
oscore
impuslive

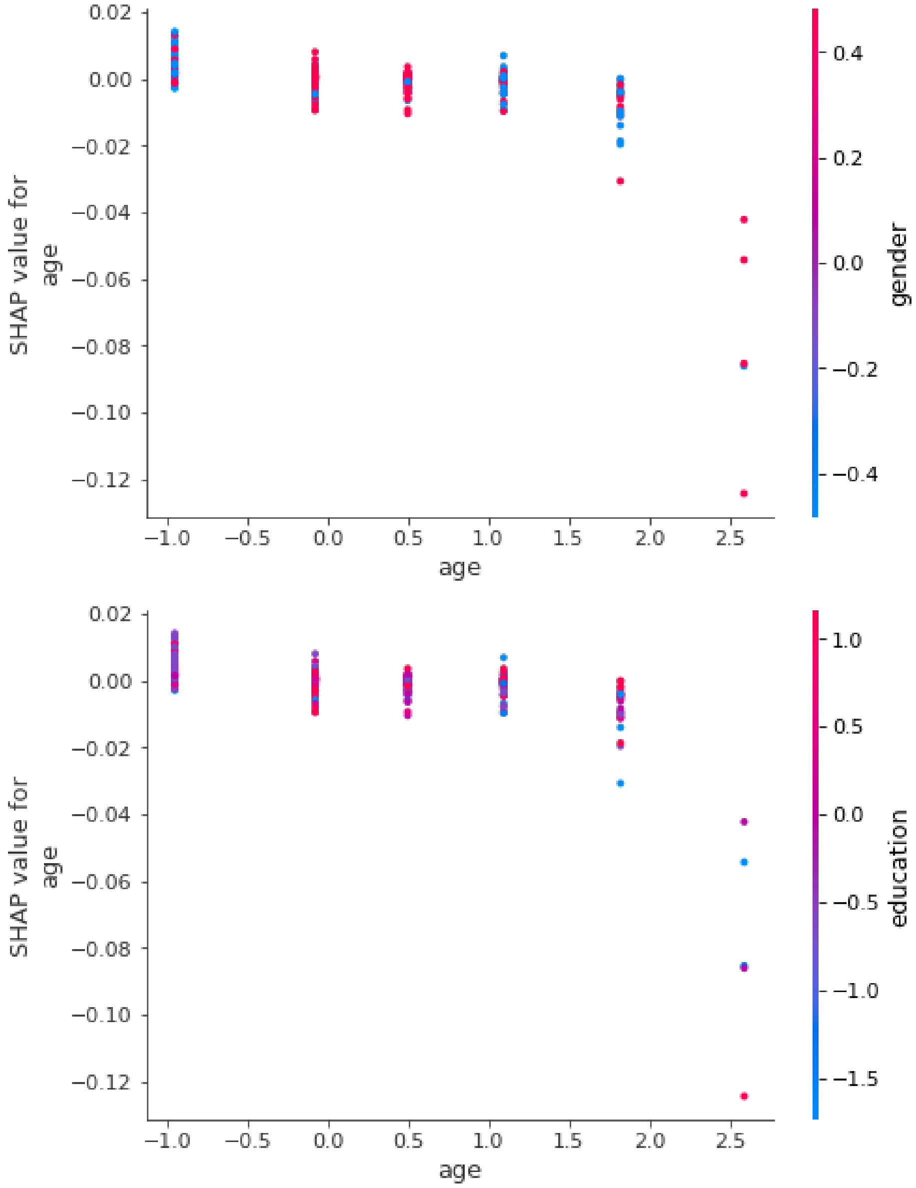

SHAP value for ascore
ascore
escore

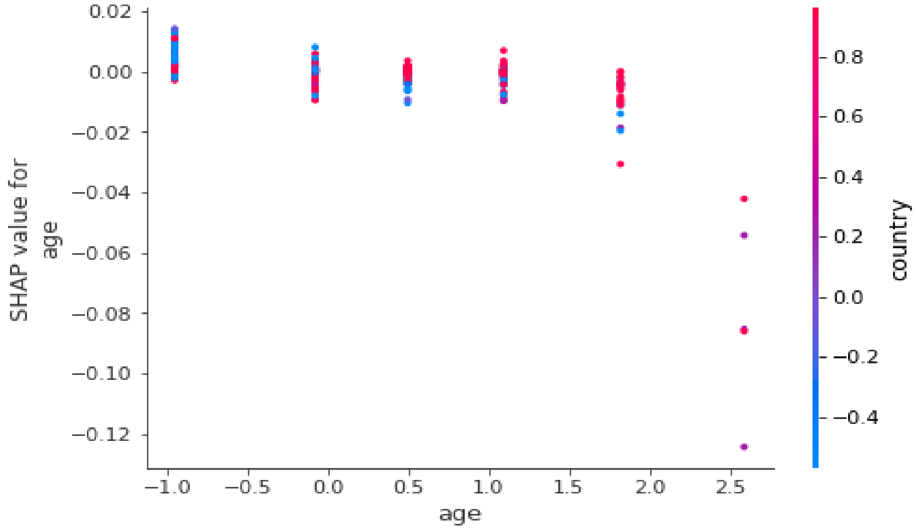

SHAP value for cscore
cscore
ethnicity
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.05
−0.06
−2
−1
0
1
2
3
−0.24
−0.26
−0.28
−0.30

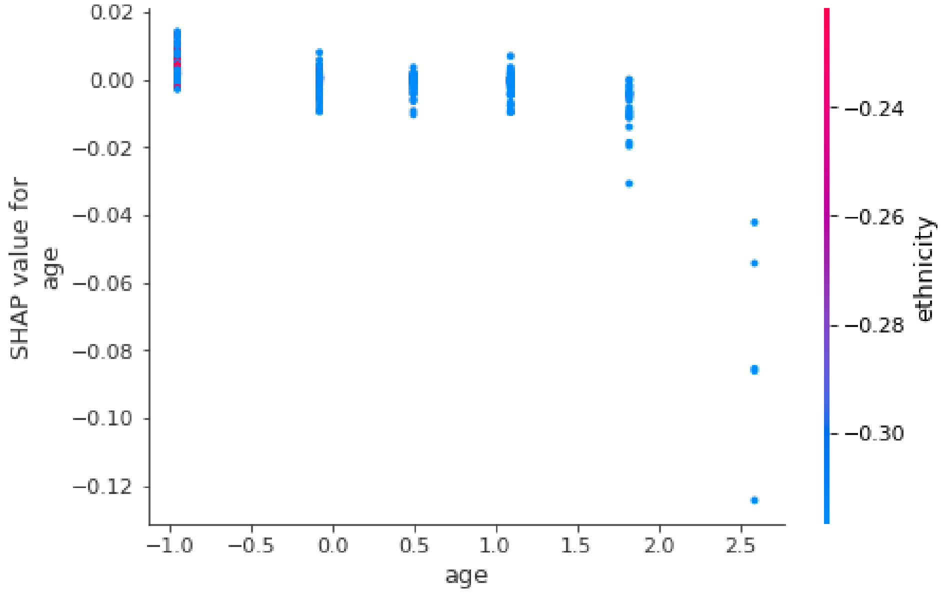

SHAP value for impuslive
impuslive
education
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
−2
−1
0
1
2
3
1.0
0.5
0.0
−0.5
−1.0
−1.5

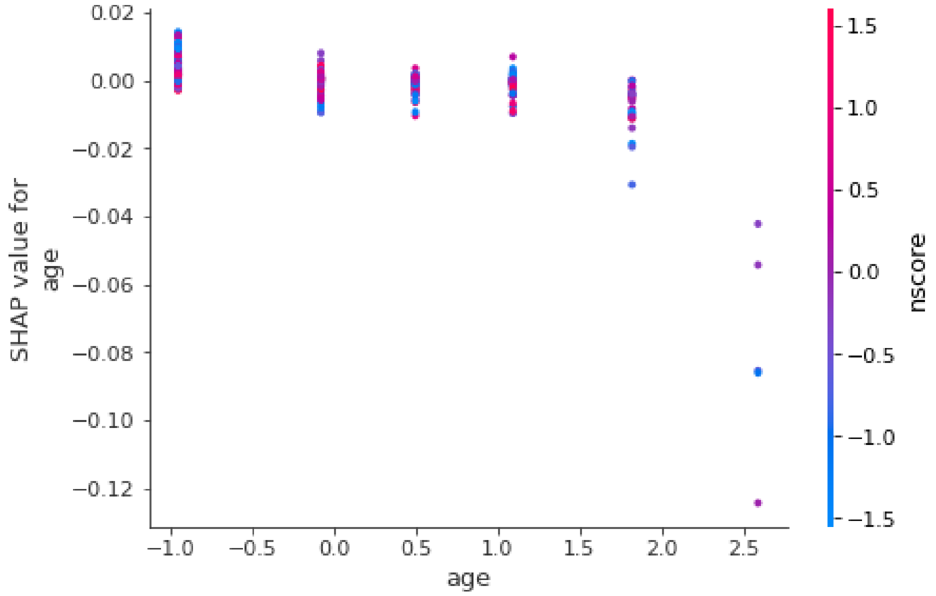

SHAP value for ss
ss
nscore
0.01
0.00
−0.01
−0.02
−0.03
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

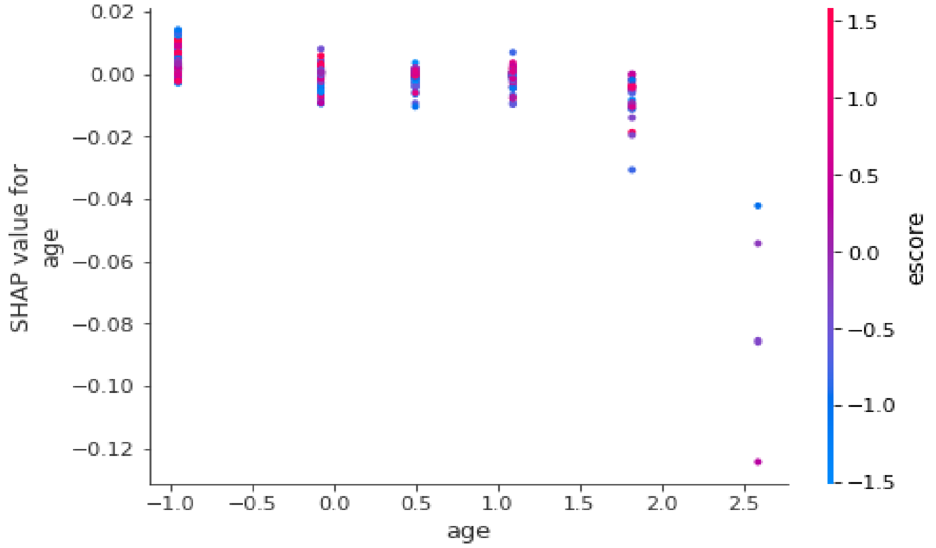
SHAP INTERACTION Plot on Model Random Forest with Dataset Chocolate-TEST

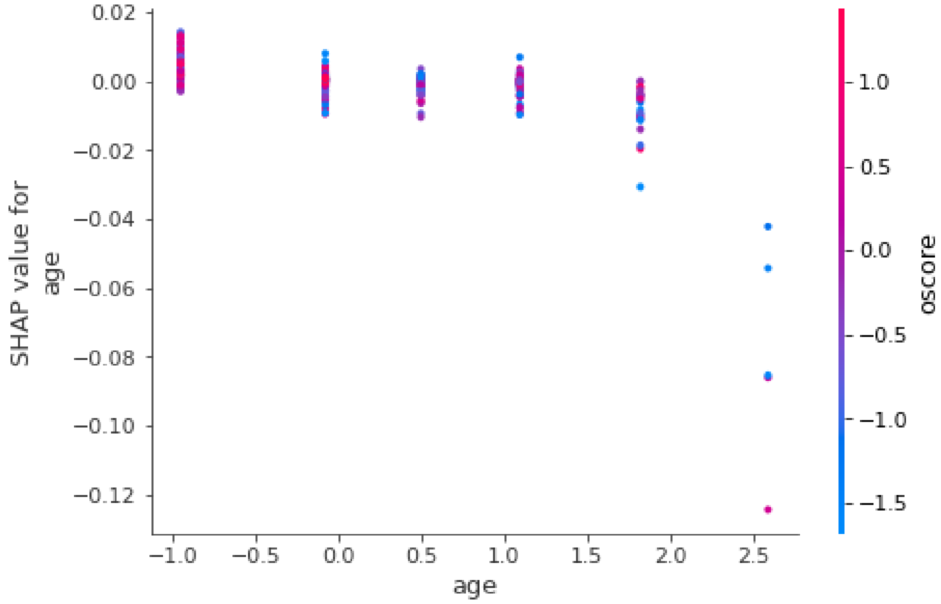


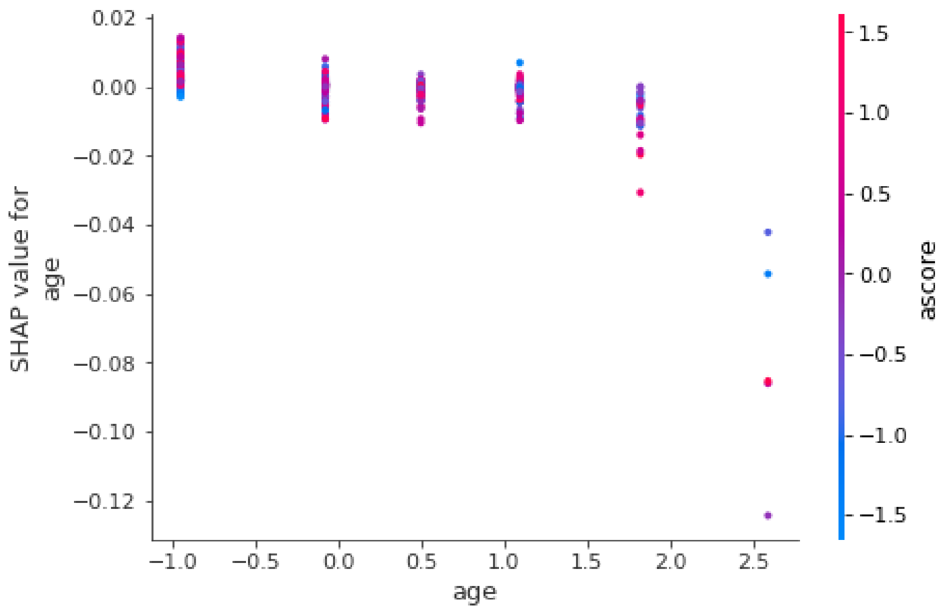

SHAP value for age
age
country
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

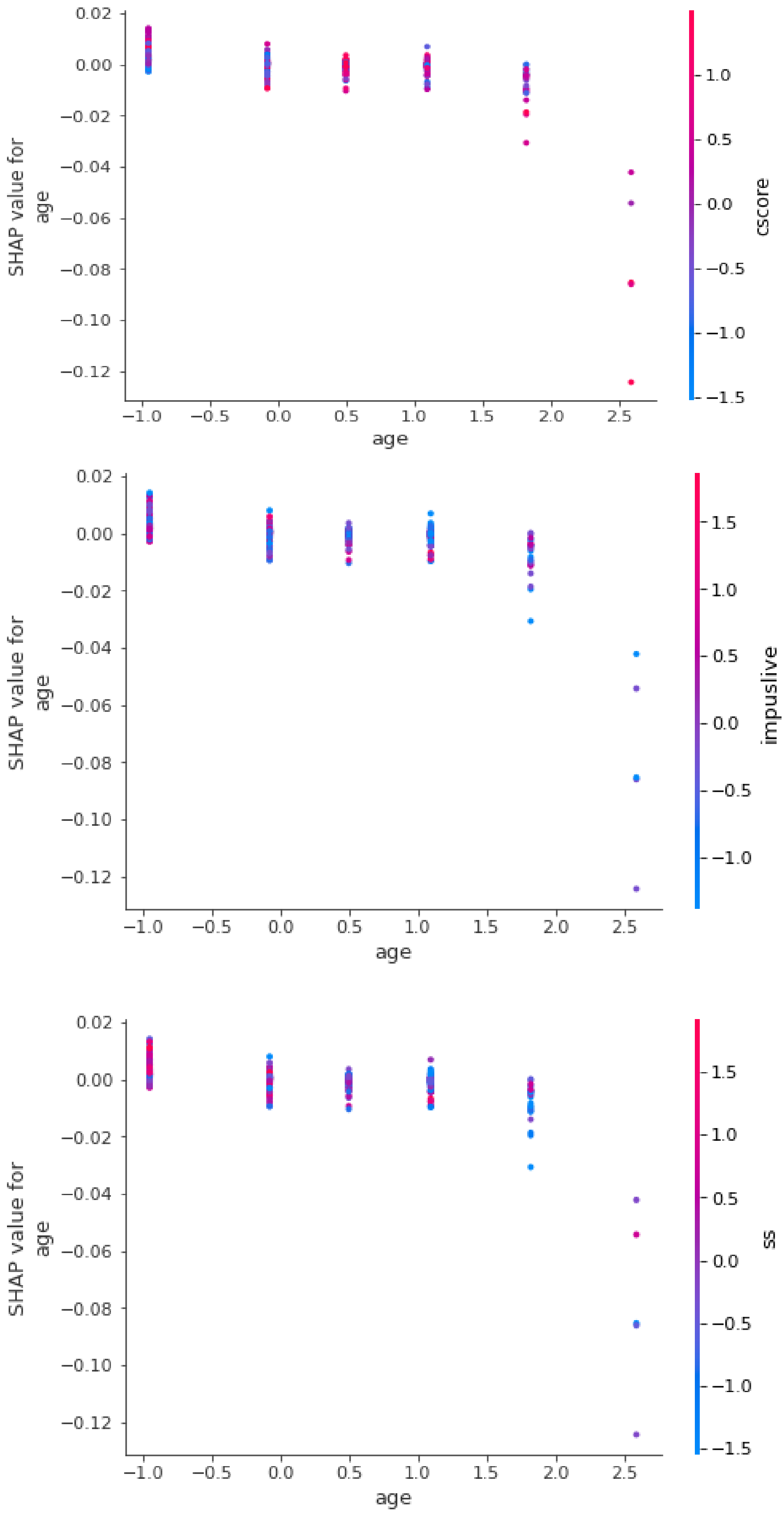

SHAP value for age
age
ethnicity
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
−0.24
−0.26
−0.28
−0.30


SHAP value for age
age
nscore
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
escore

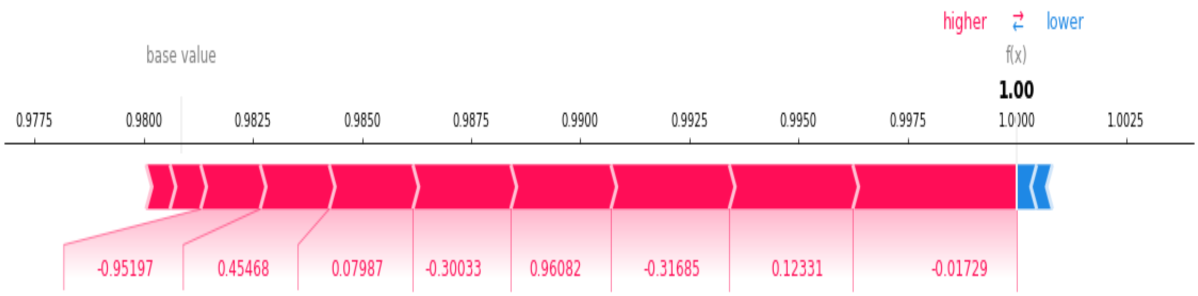

0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
1.0
0.5
0.0
−0.5
−1.0
−1.5
oscore


0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
ascore

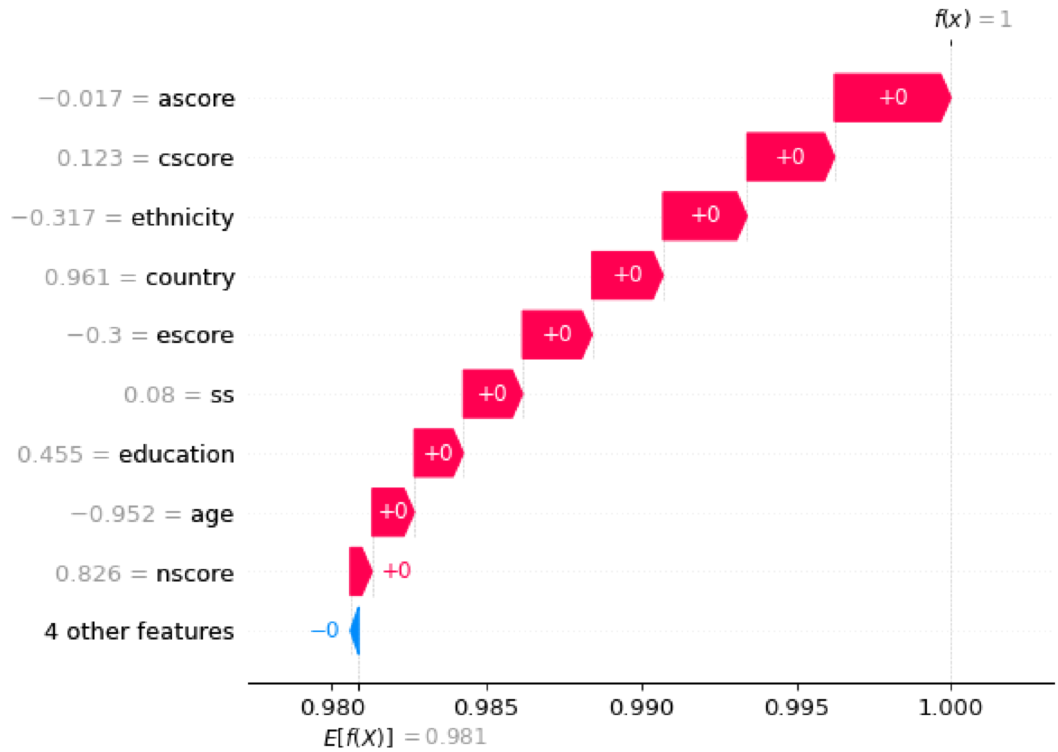


SHAP value for age
age
cscore
impuslive
ss

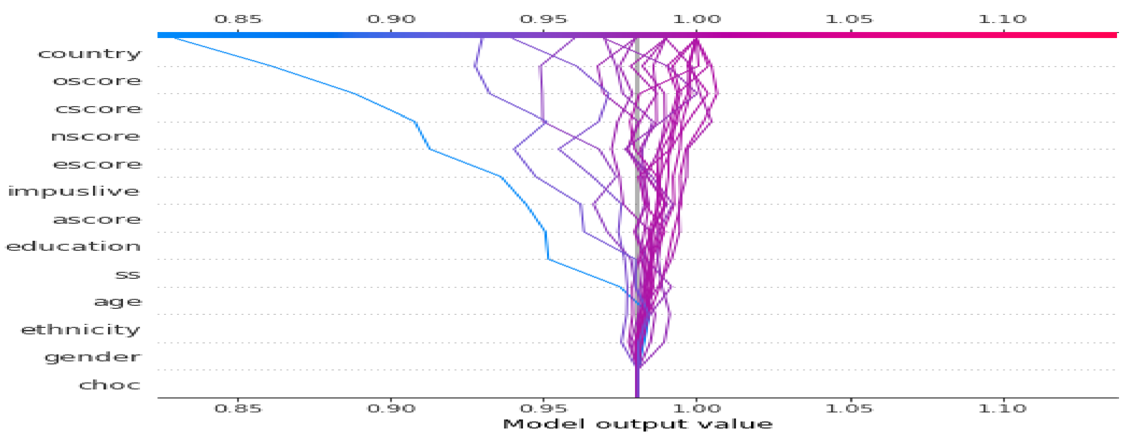
SHAP Force Plot for a single prediction in Model Random Forest



SHAP Waterfall Plot for a single prediction in Model Random Forest

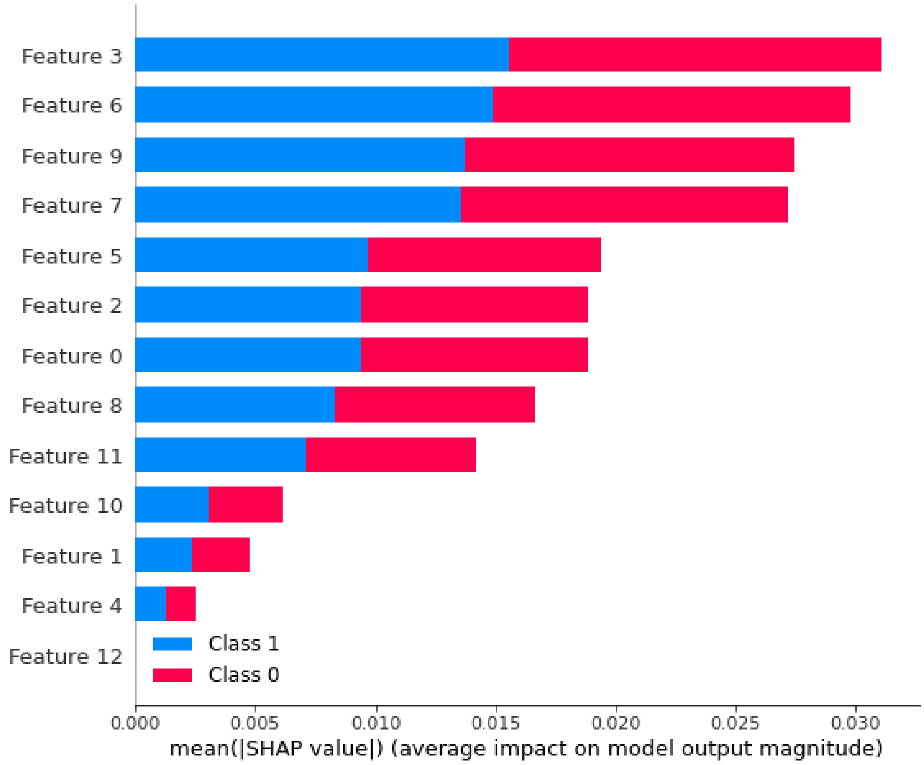


SHAP Decision Plot for Model Random Forest on Dataset Chocolate-TEST

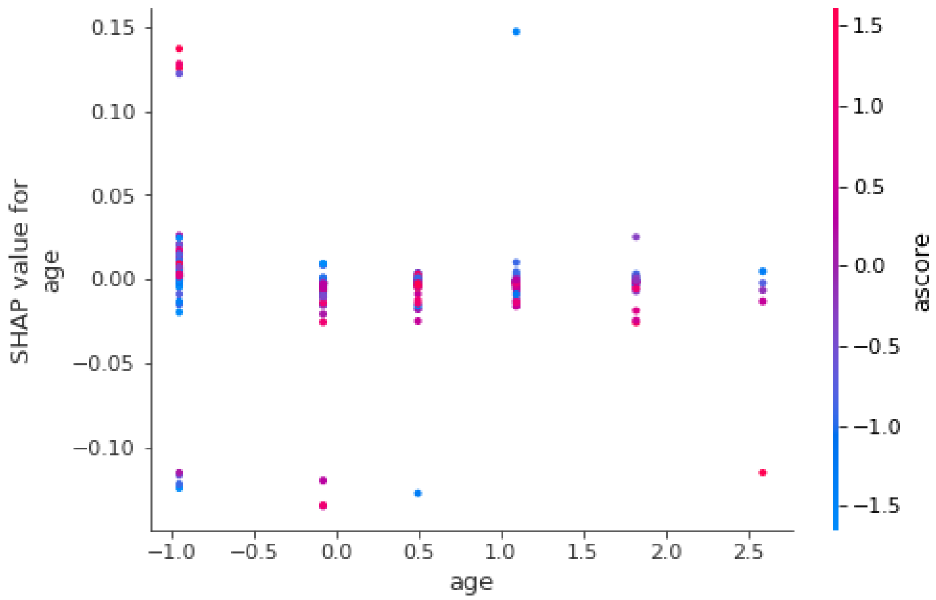


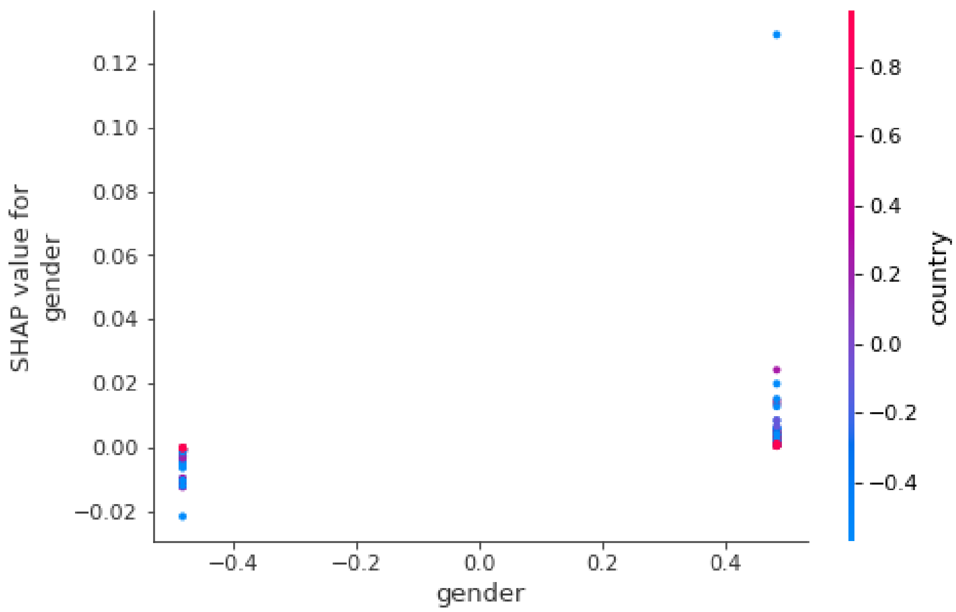
Computing SHAP values with the worst model (Decision Tree) on Chocolate test set

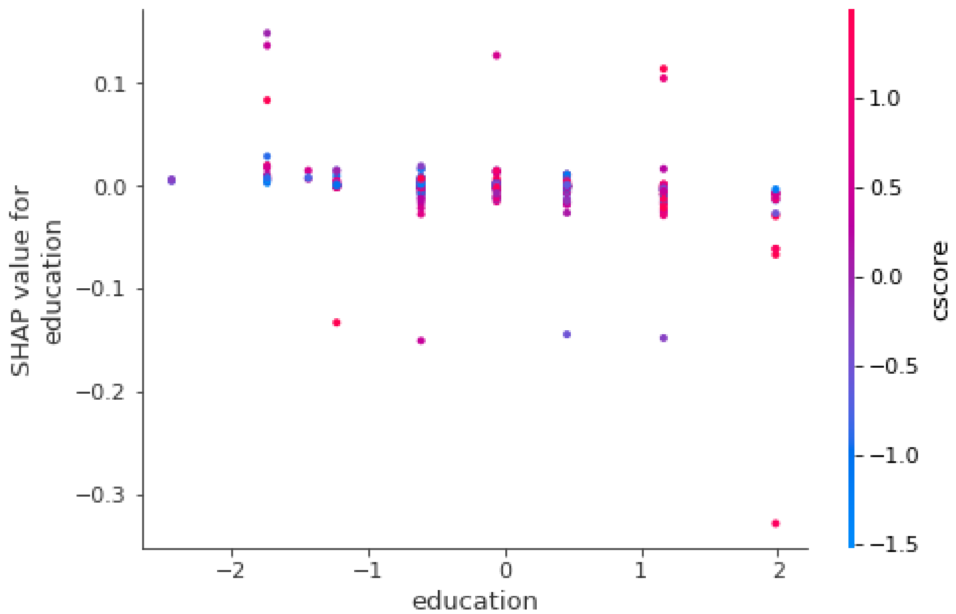
```
SHAP Summary Plot for Model Decision Tree on Dataset Chocolate Test Set
```

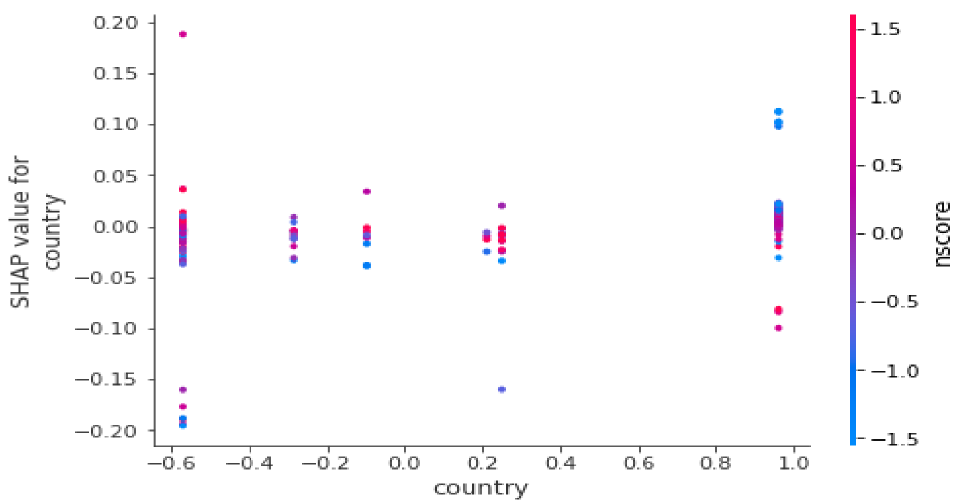


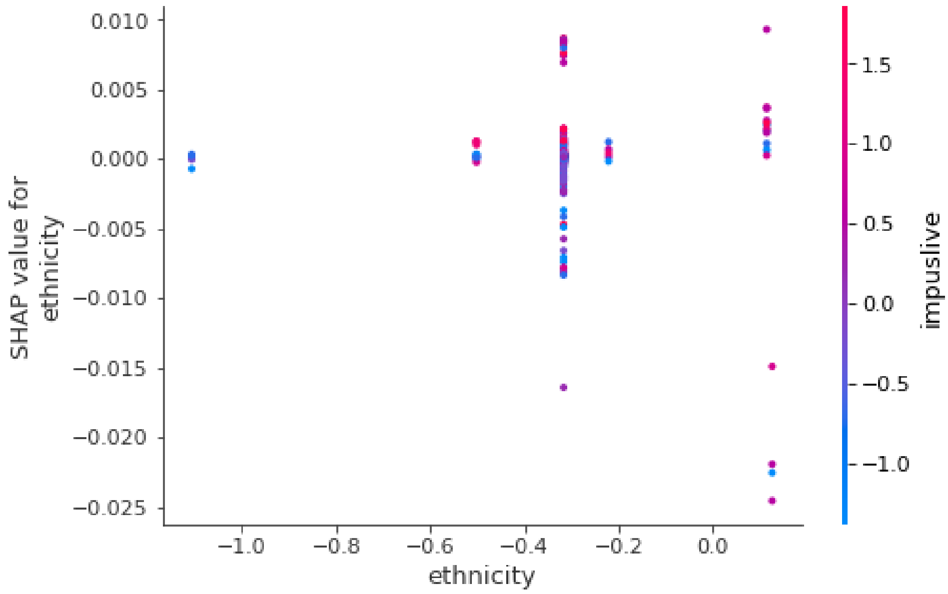

```
SHAP Dependence Plot on Model Decision Tree with Dataset Chocolate Test
Set
```

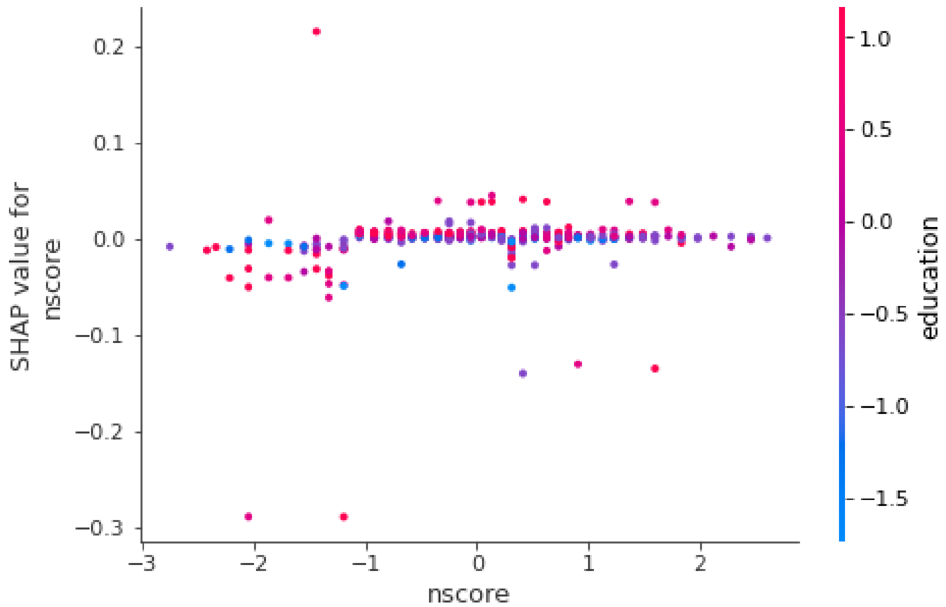


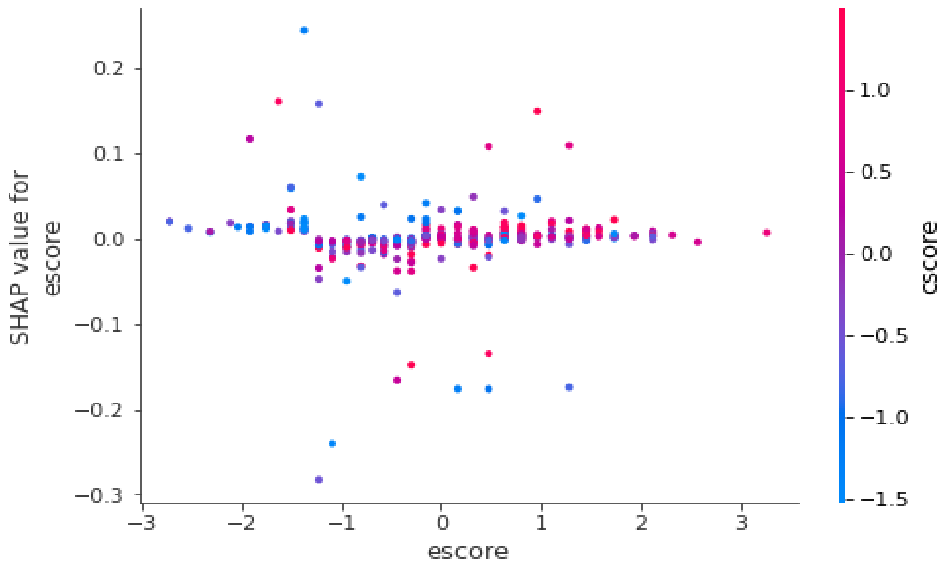

SHAP value for gender
gender
country
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
−0.4
−0.2
0.0
0.2
0.4
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

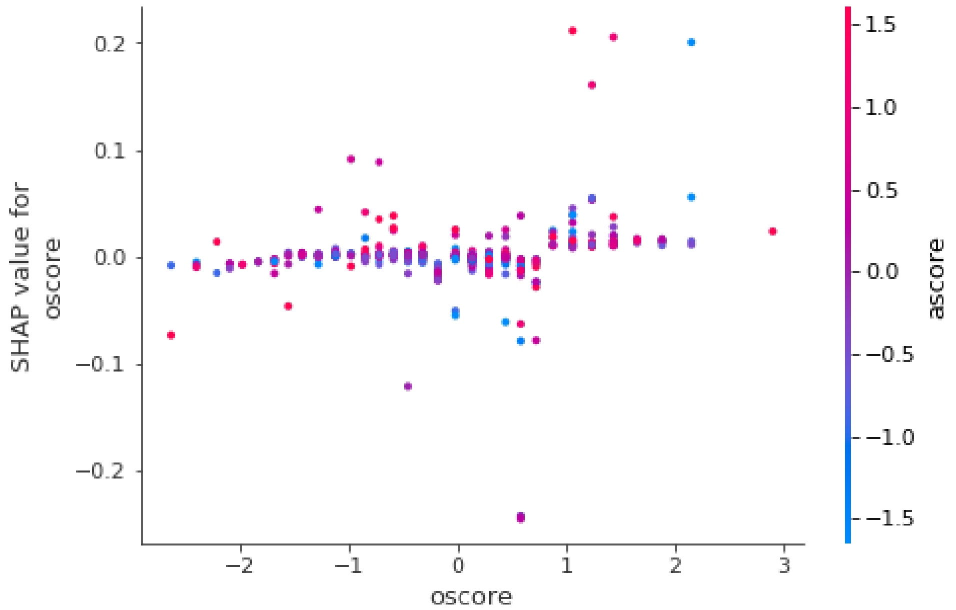

SHAP value for education
education
cscore
0.1
0.0
−0.1
−0.2
−0.3
−2
−1
0
1
2
1.0
0.5
0.0
−0.5
−1.0
−1.5

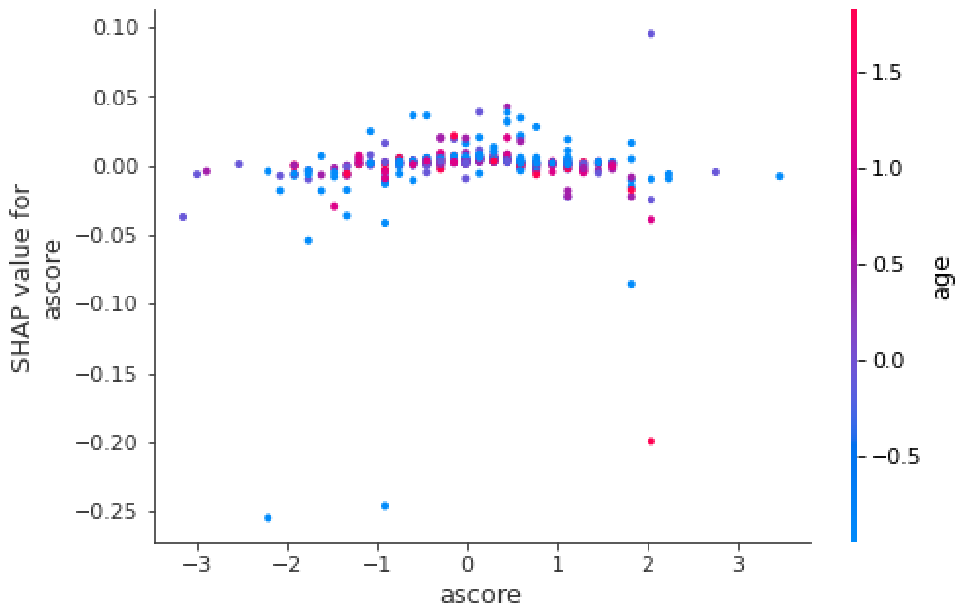

SHAP value for country
country
nscore
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

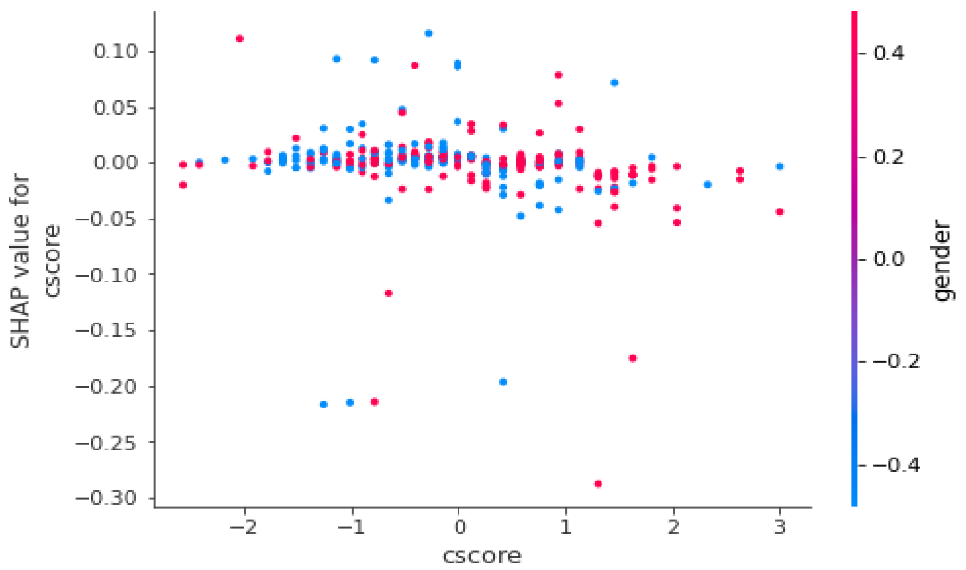

SHAP value for ethnicity
ethnicity
impuslive

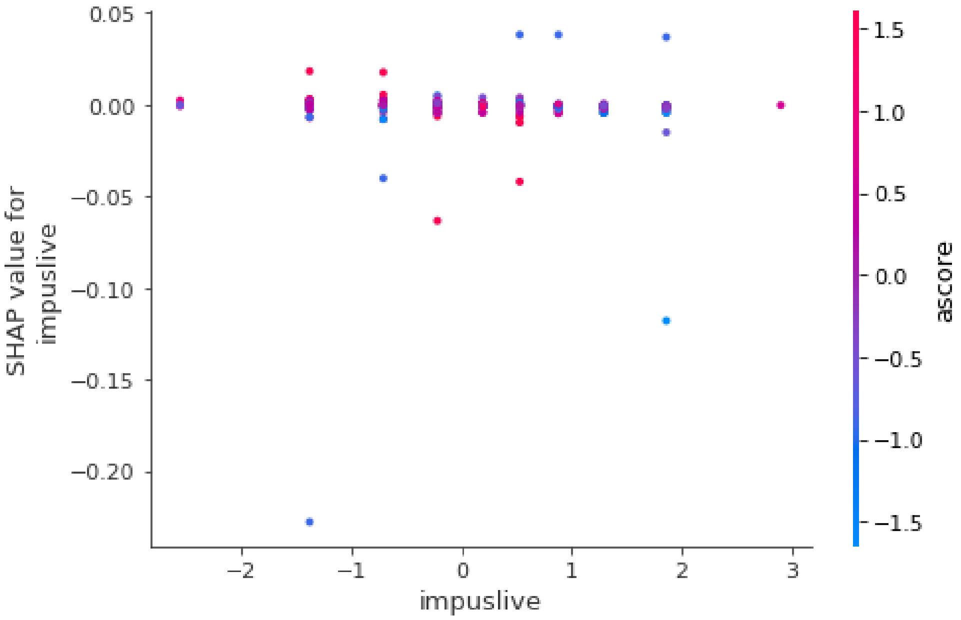

SHAP value for nscore
nscore
education

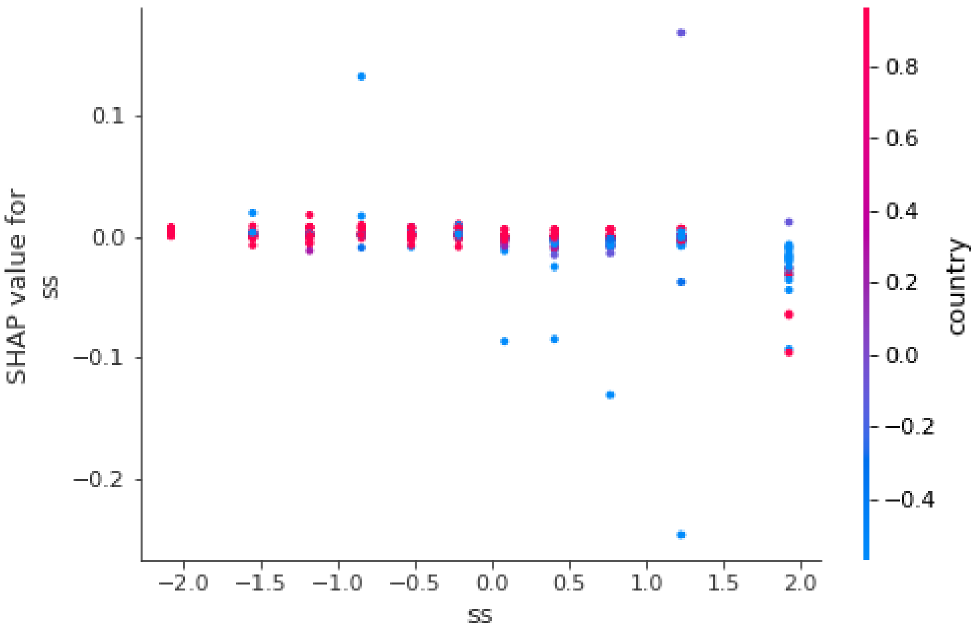

SHAP value for escore
escore
cscore


SHAP value for oscore
oscore
ascore
0.2
0.1
0.0
−0.1
−0.2
−2
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

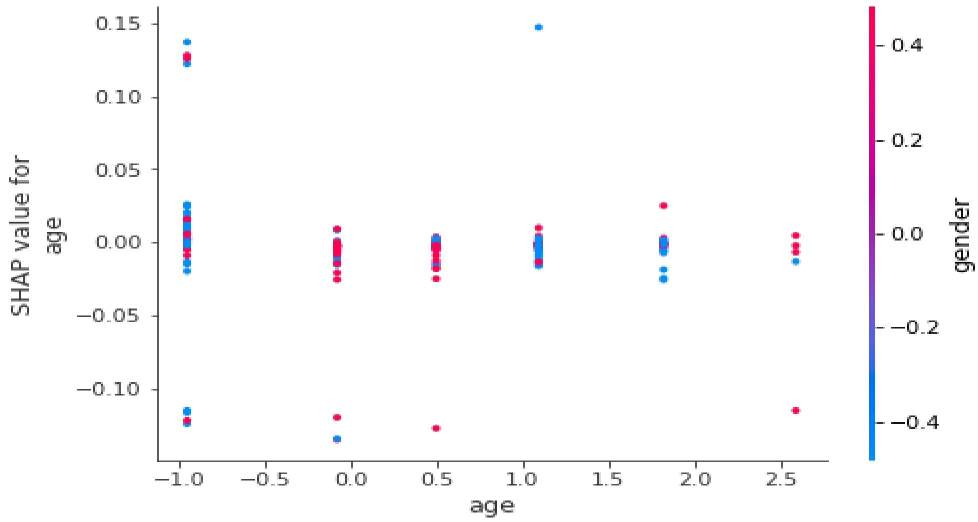

SHAP value for ascore
ascore
age
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−3
−2
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5

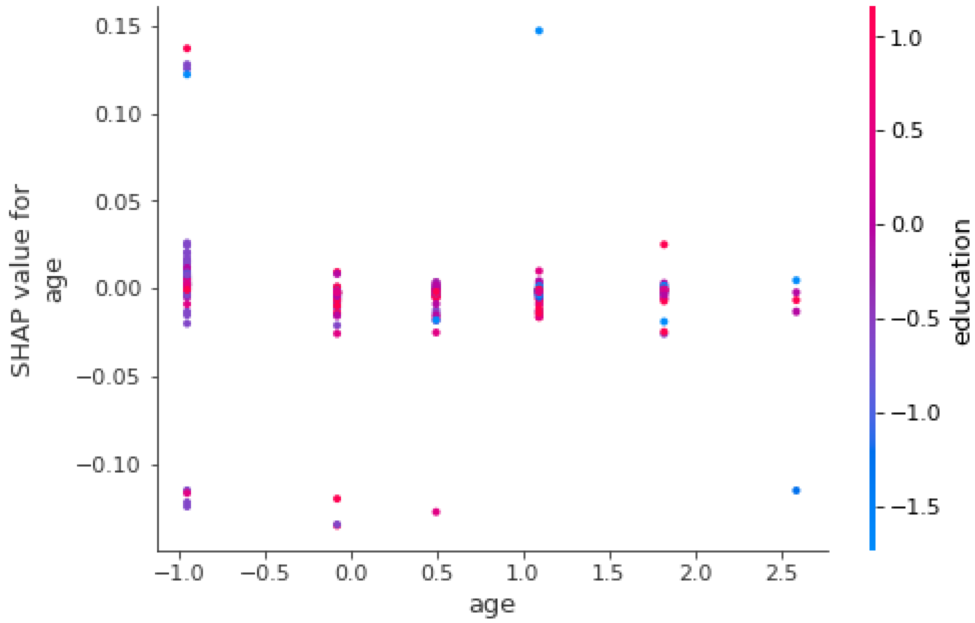

SHAP value for cscore
cscore
gender
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
−2
−1
0
1
2
3
0.4
0.2
0.0
−0.2
−0.4

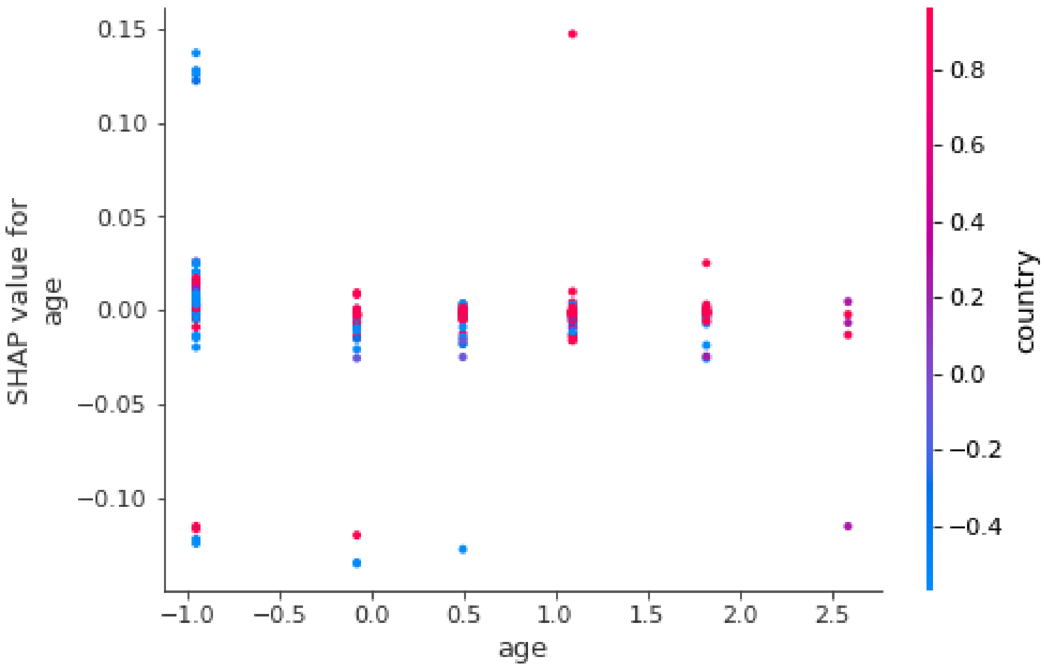

0.05
0.00
−0.05
−0.10
−0.15
−0.20
SHAP value for impuslive
−2
−1
0
1
2
3
impuslive
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
ascore

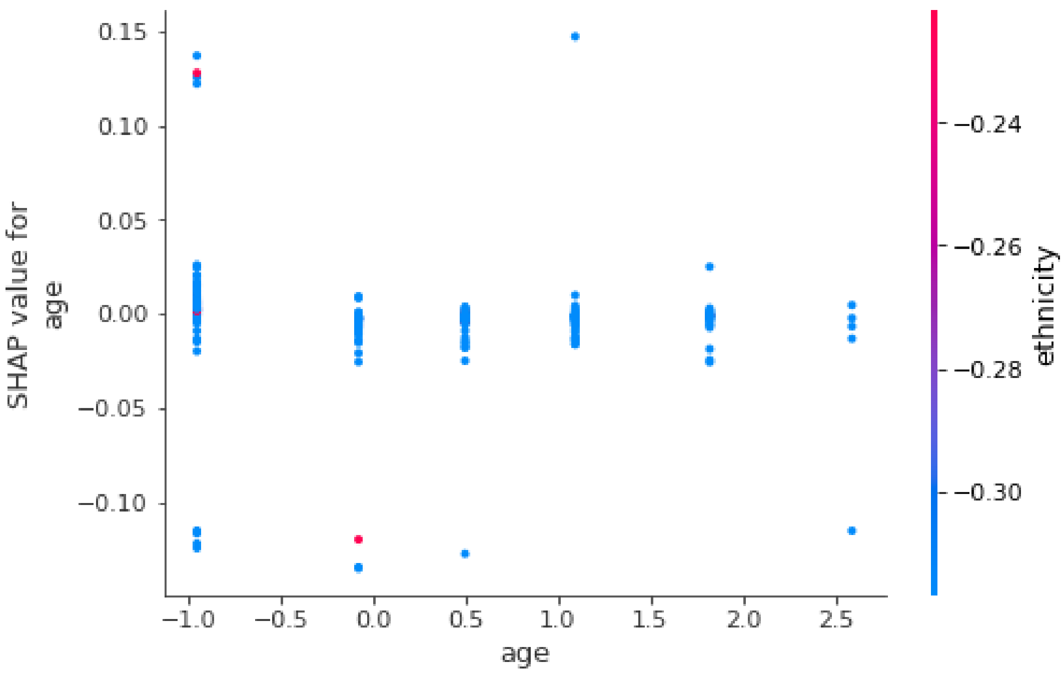

0.1
0.0
−0.1
−0.2
SHAP value for ss
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
ss
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
country

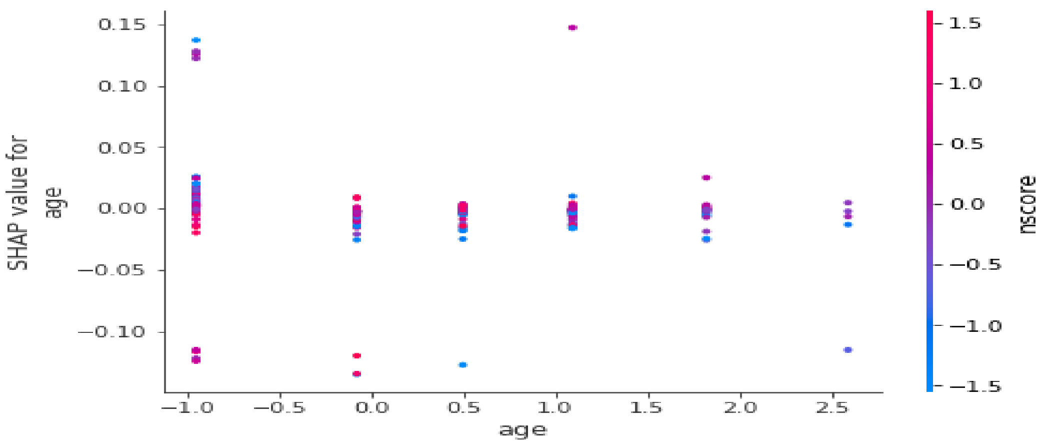
SHAP INTERSECTION Plot on Model Decision Tree with Dataset Chocolate Test Set

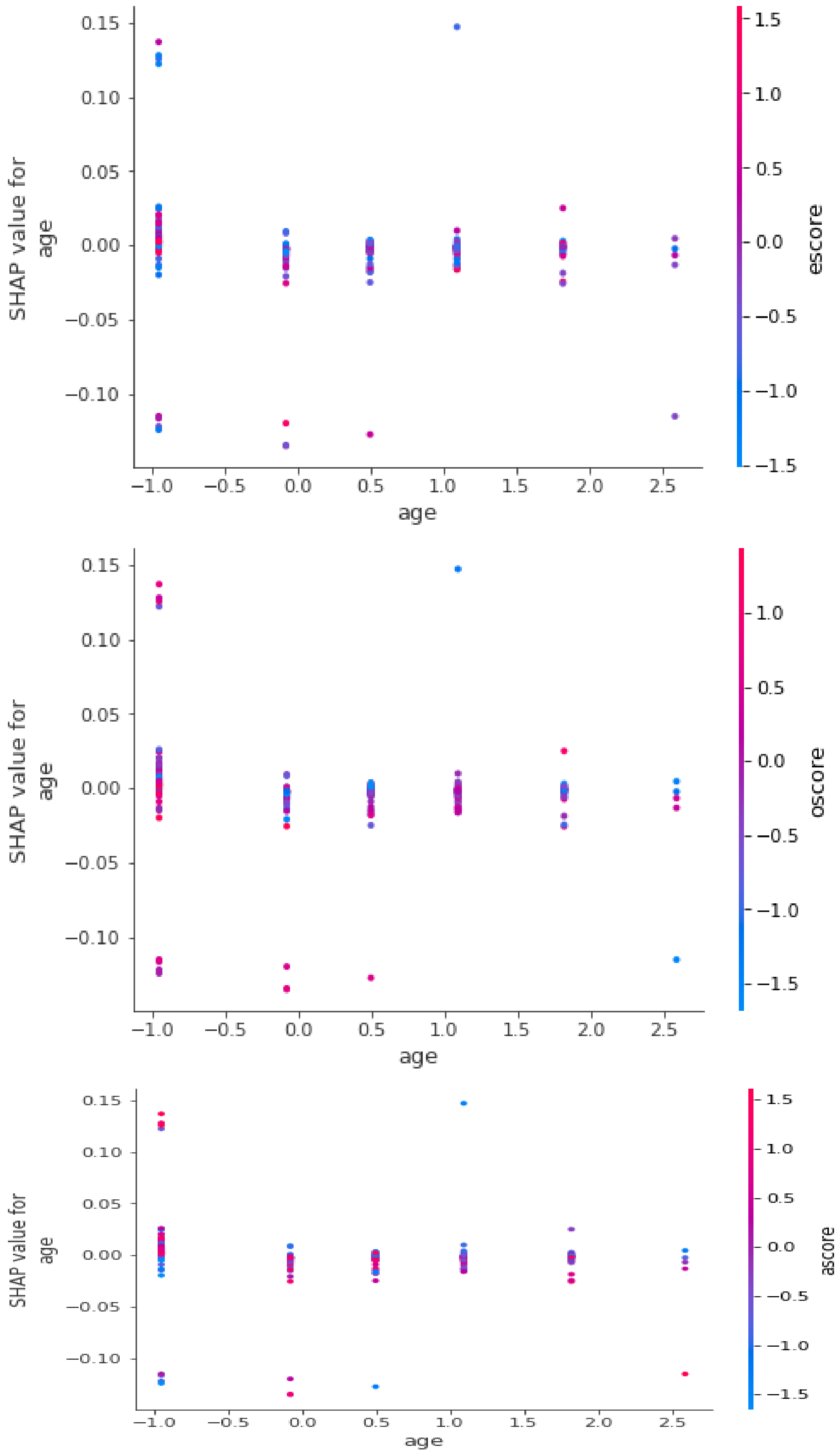


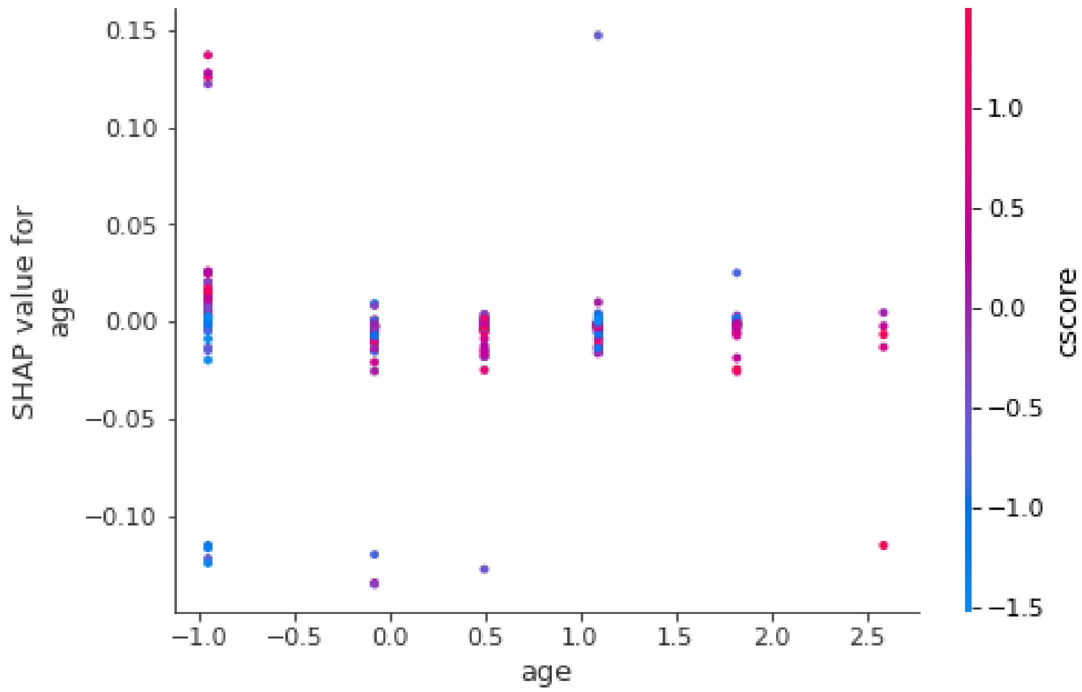


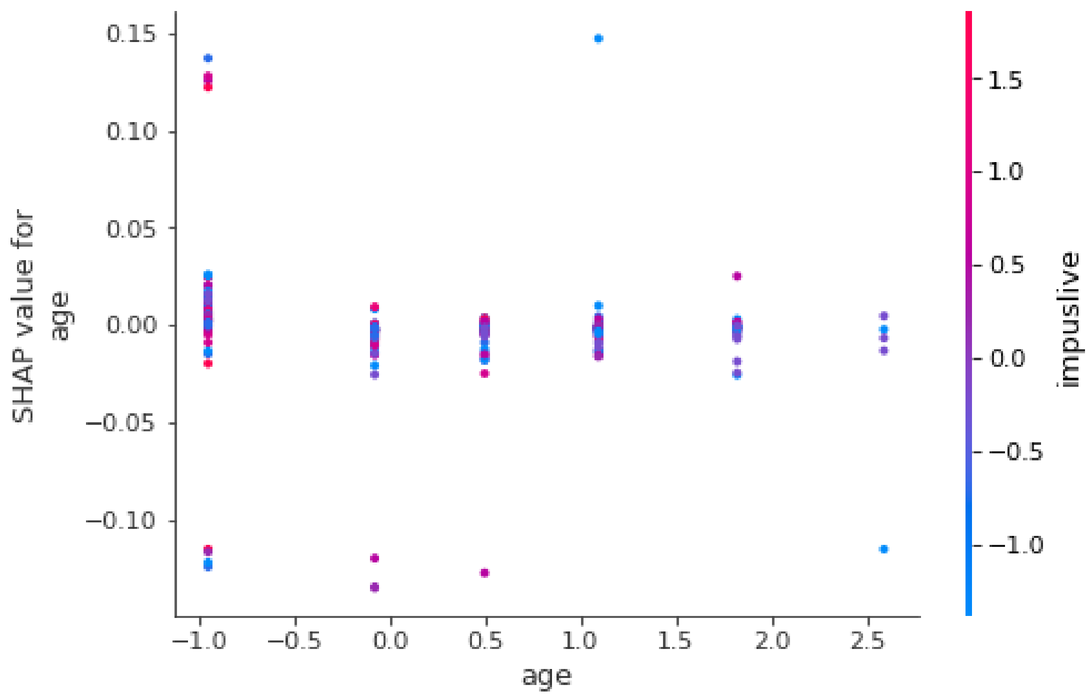

0.15
0.10
0.05
0.00
−0.05
−0.10
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
country

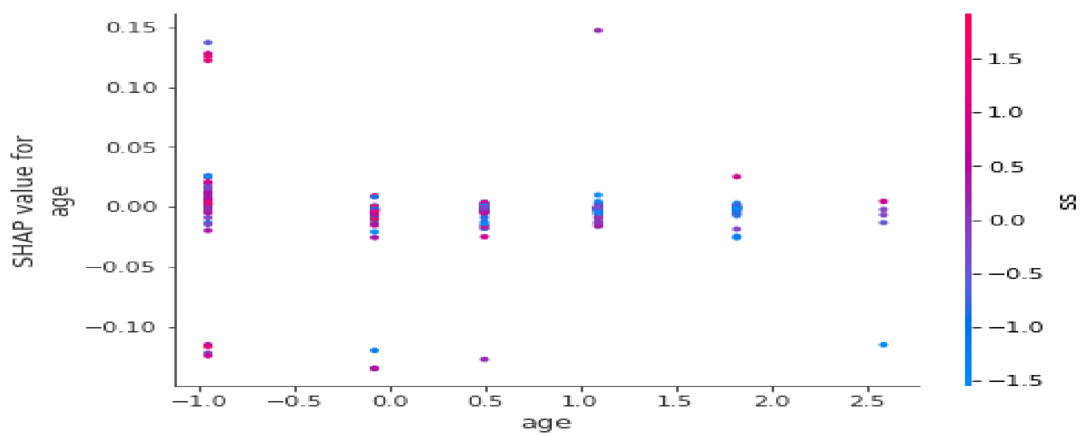

0.15
0.10
0.05
0.00
−0.05
−0.10
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
−0.24
−0.26
−0.28
−0.30
ethnicity


0.15
0.10
0.05
0.00
−0.05
−0.10
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
nscore

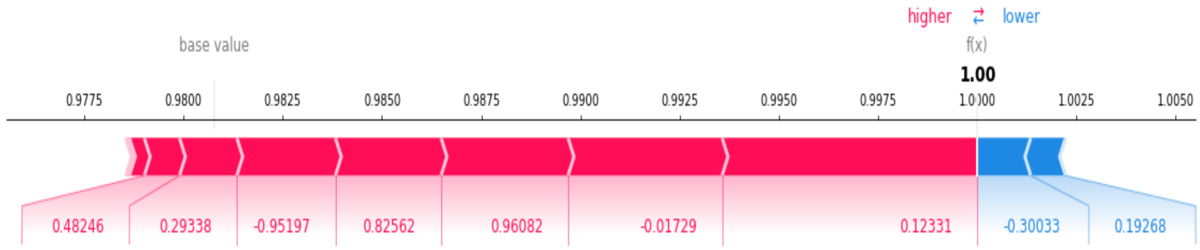

SHAP value for age
age
escore
oscore
ascore


SHAP value for age
age
cscore

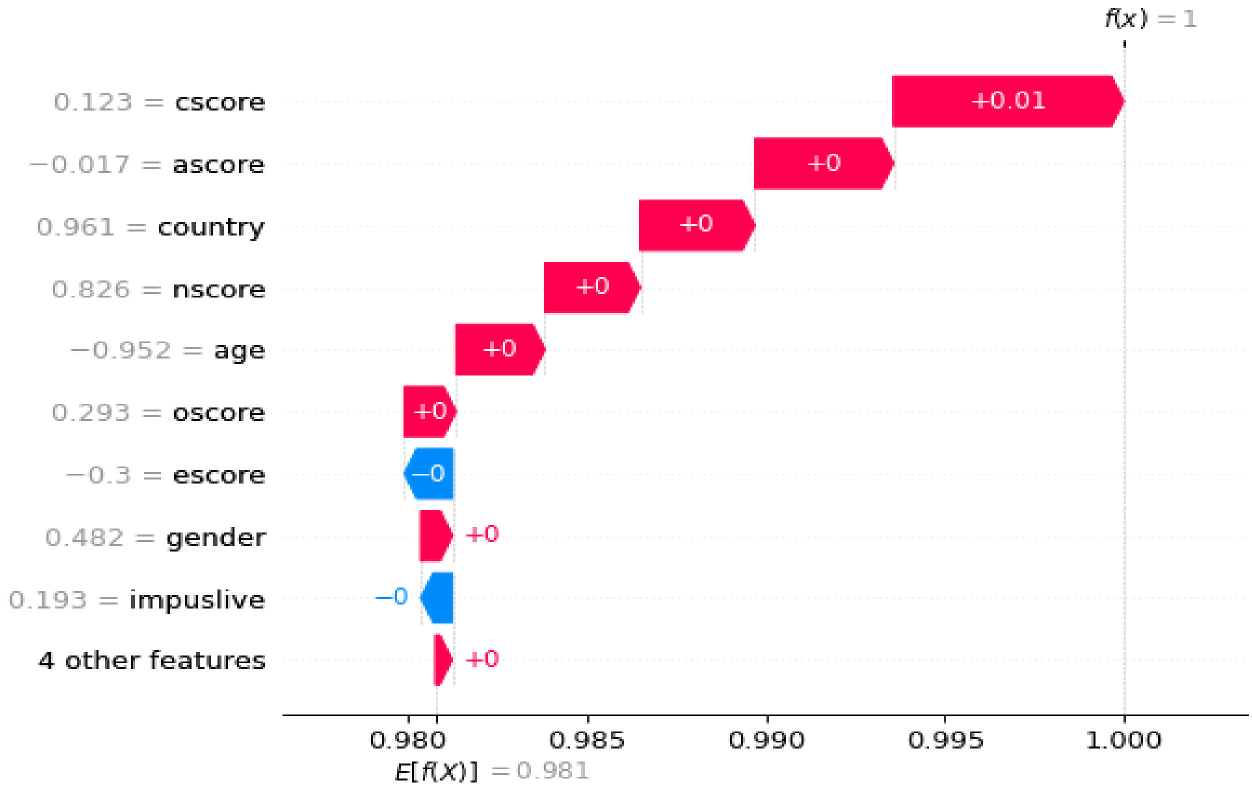

SHAP value for age
age
impuslive


SHAP value for age
age
ss

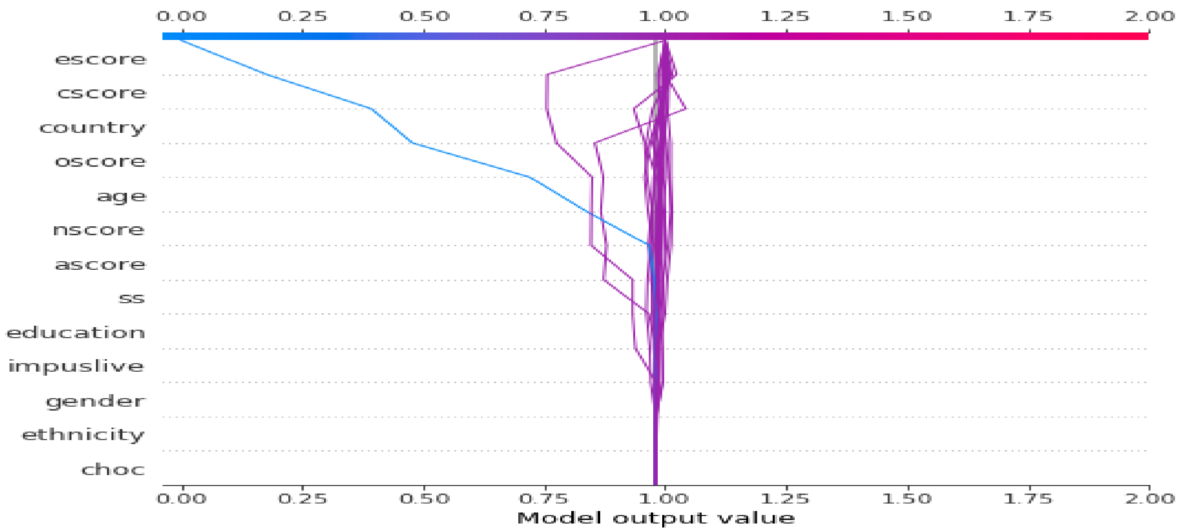
SHAP Force Plot for a single prediction in Model Decision Tree



SHAP Waterfall Plot for a single prediction in Model Decision Tree

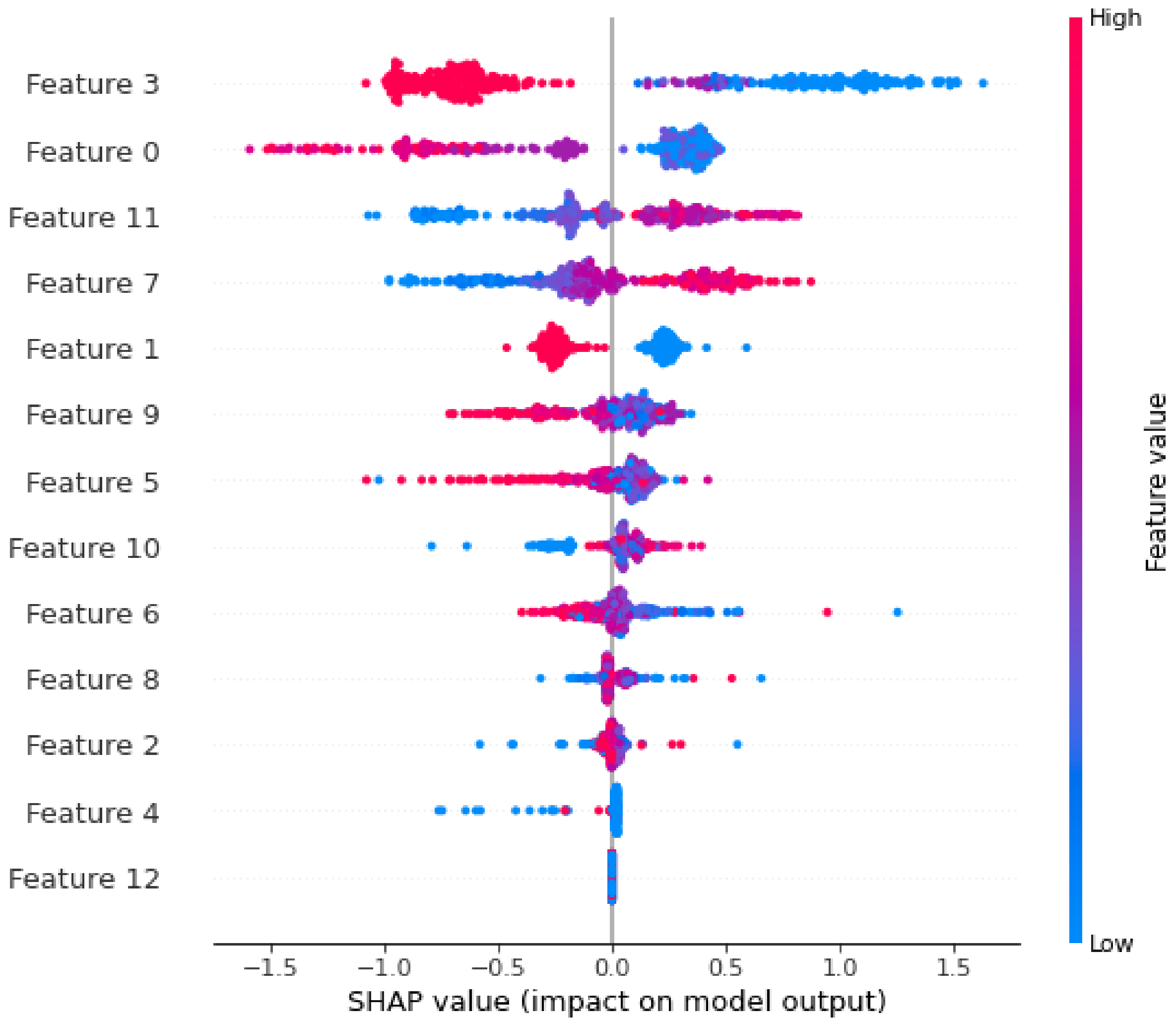


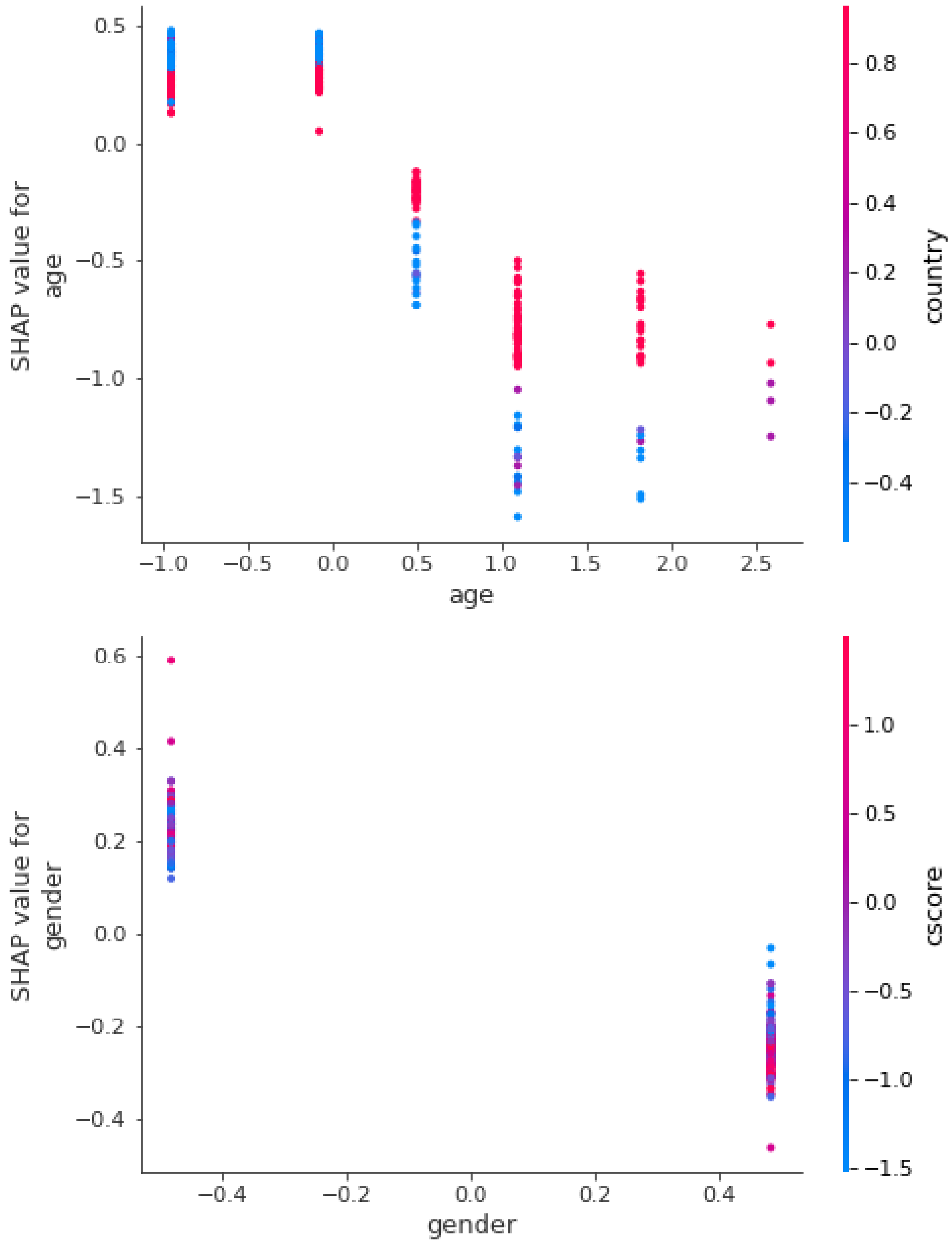
SHAP Decision Plot for Model Decision Tree on Dataset Chocolate-test



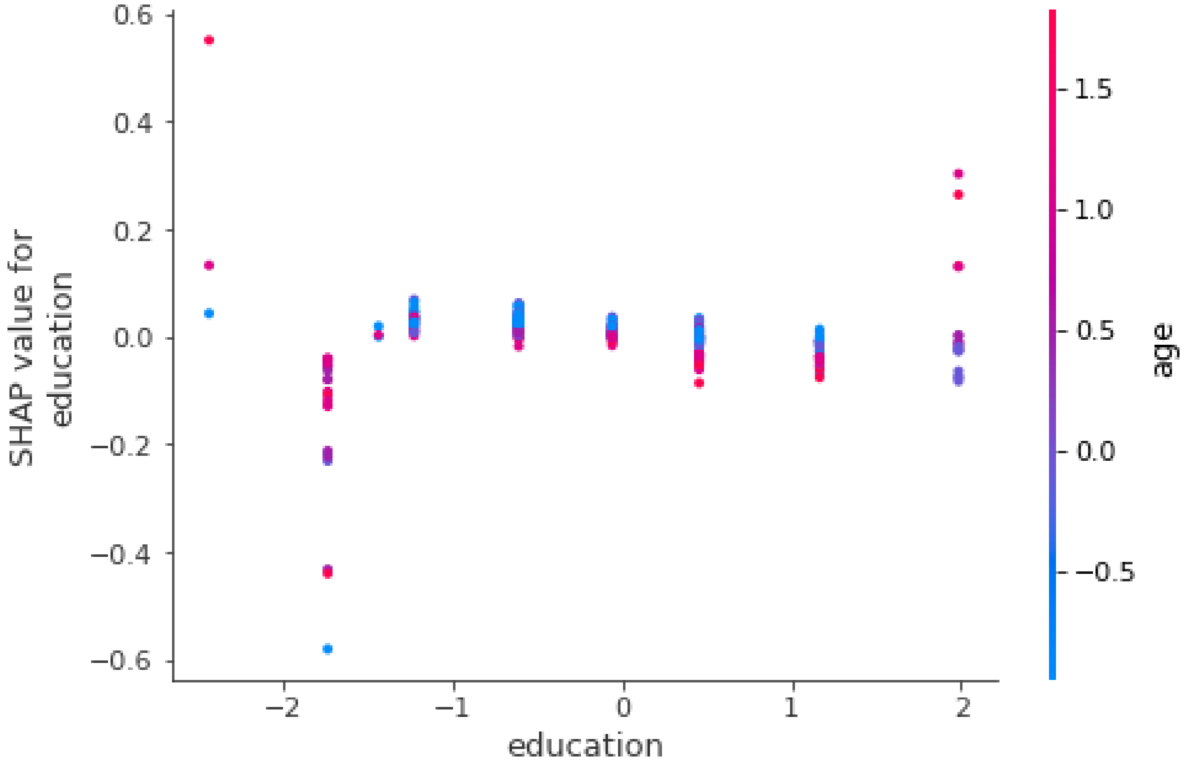
Computing SHAP values with the best model (Gradient Boost) on MUSHROOM test set

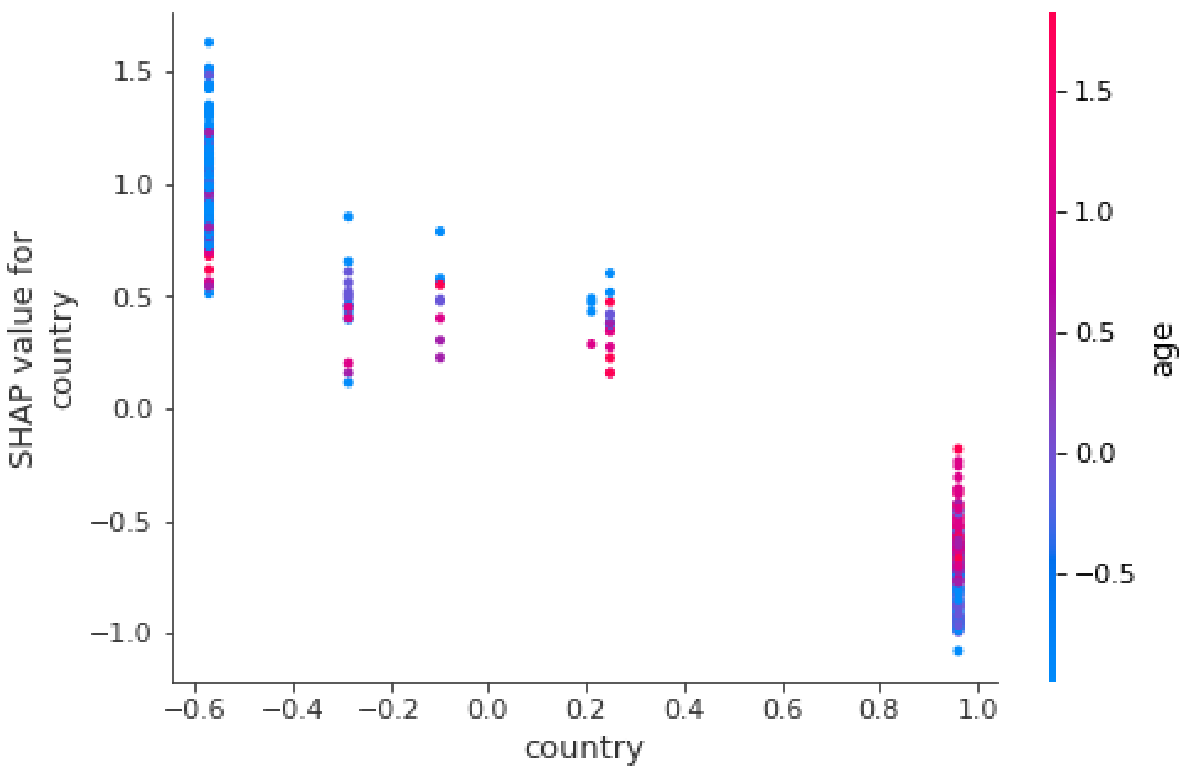
```
SHAP Summary Plot for Model GradientBoost on Dataset Mushroom Test Set
```

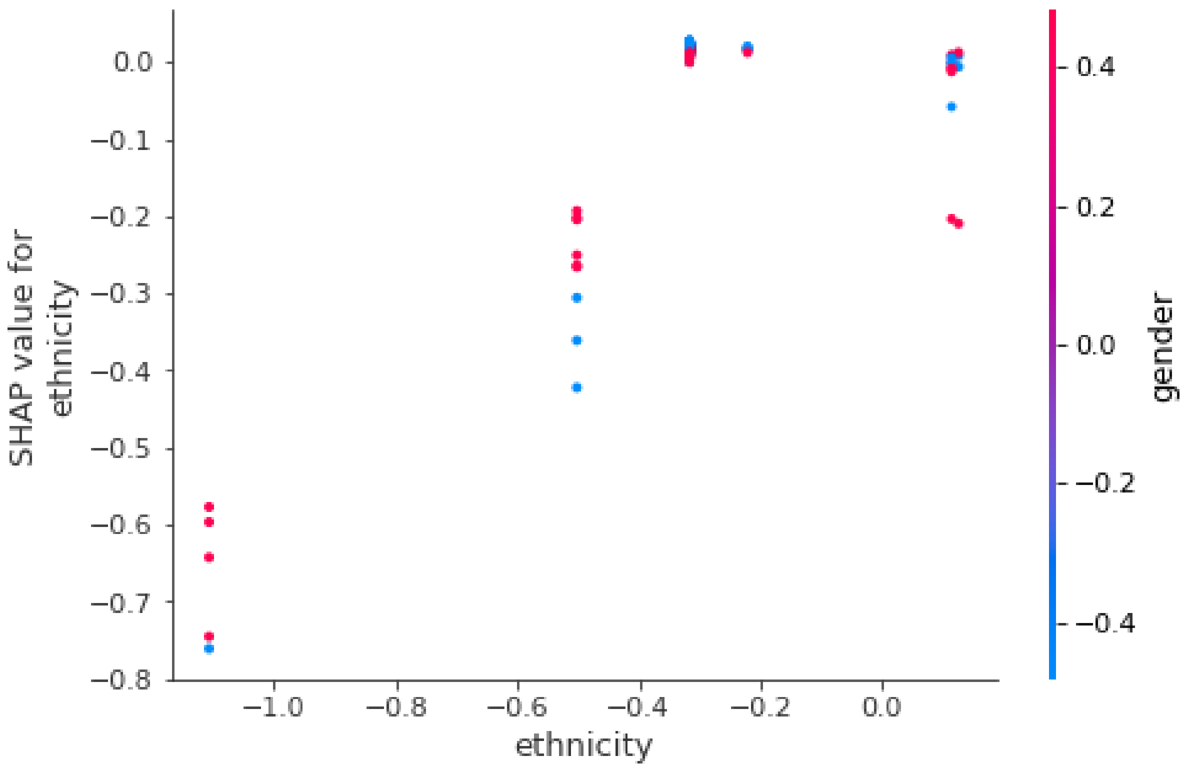


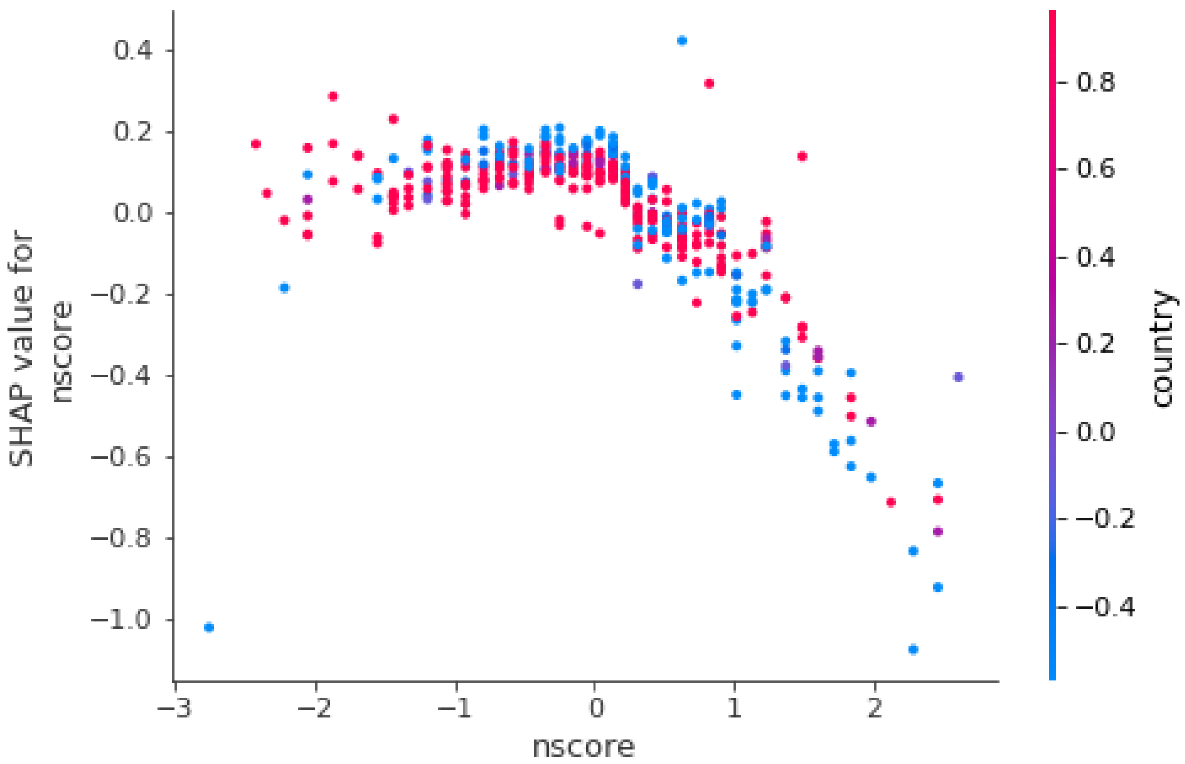


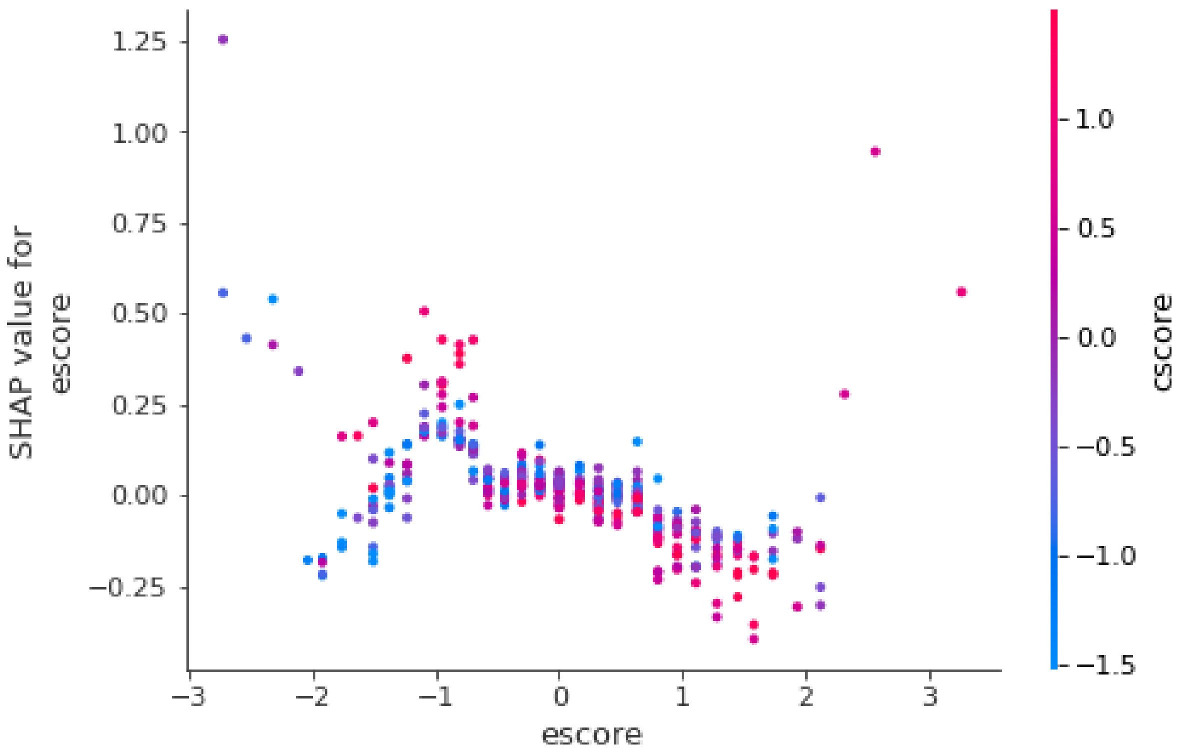
SHAP value for age
age
country
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
SHAP value for gender
gender
cscore
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
1.0
0.5
0.0
−0.5
−1.0
−1.5

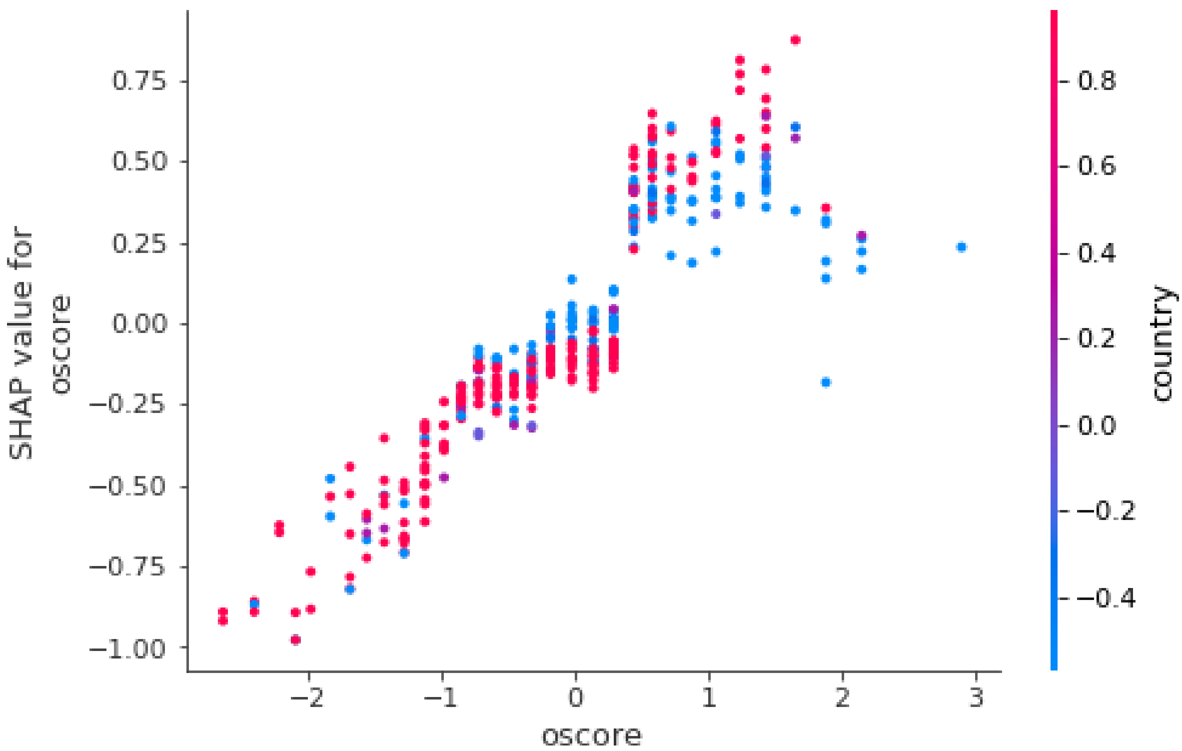

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
SHAP value for education
−2
−1
0
1
2
education
1.5
1.0
0.5
0.0
−0.5
age

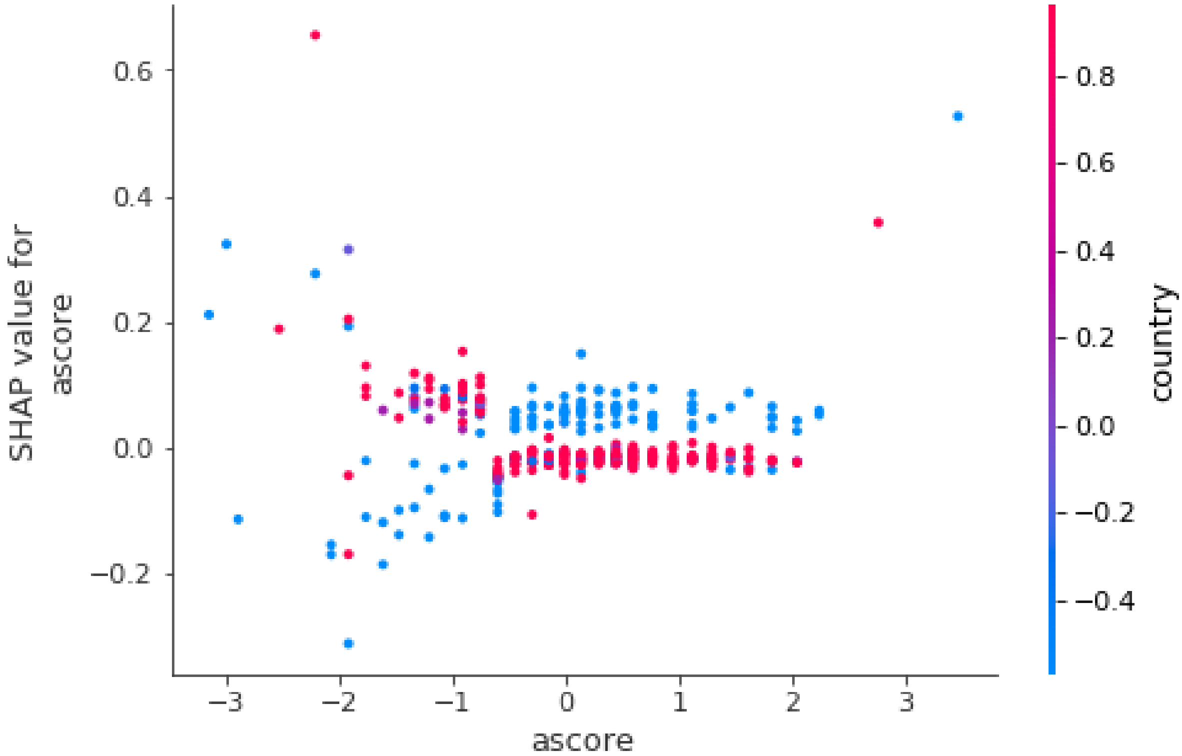

1.5
1.0
0.5
0.0
−0.5
−1.0
SHAP value for country
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
country
1.5
1.0
0.5
0.0
−0.5
age


SHAP value for
ethnicity
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
ethnicity
gender
0.4
0.2
0.0
−0.2
−0.4

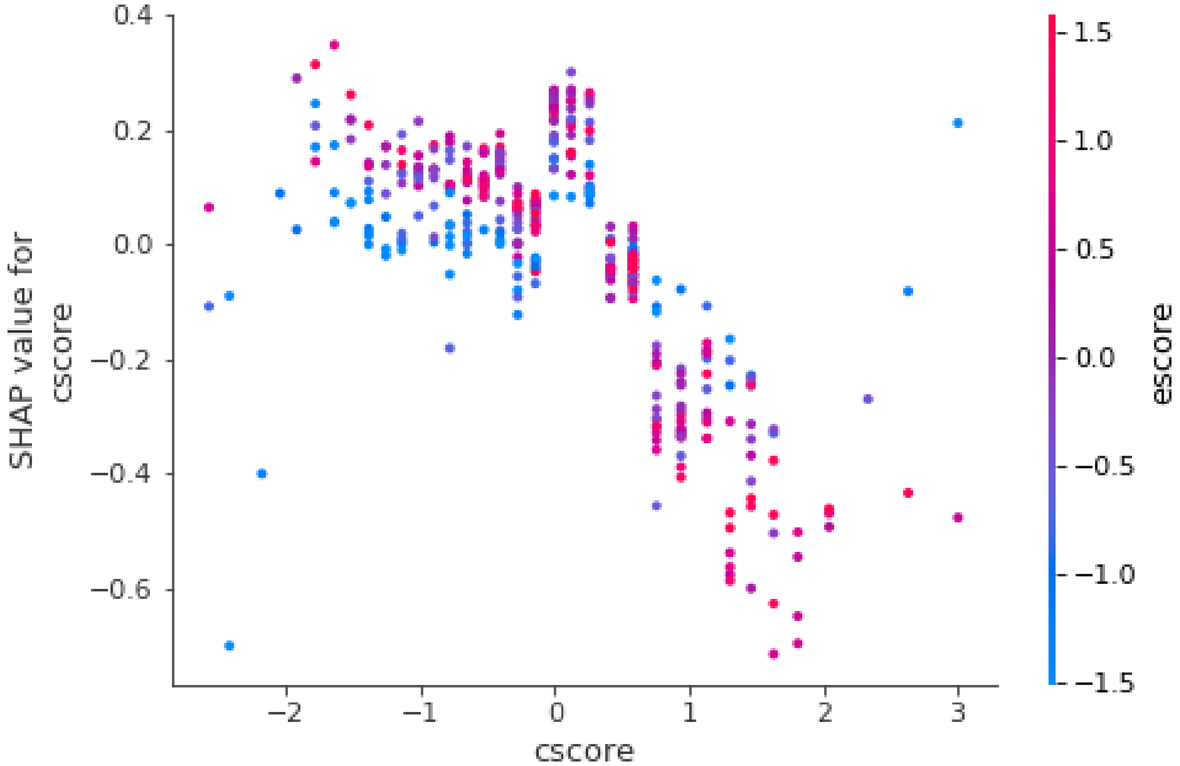

SHAP value for
nscore
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−3
−2
−1
0
1
2
nscore
country
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


SHAP value for escore
escore
cscore

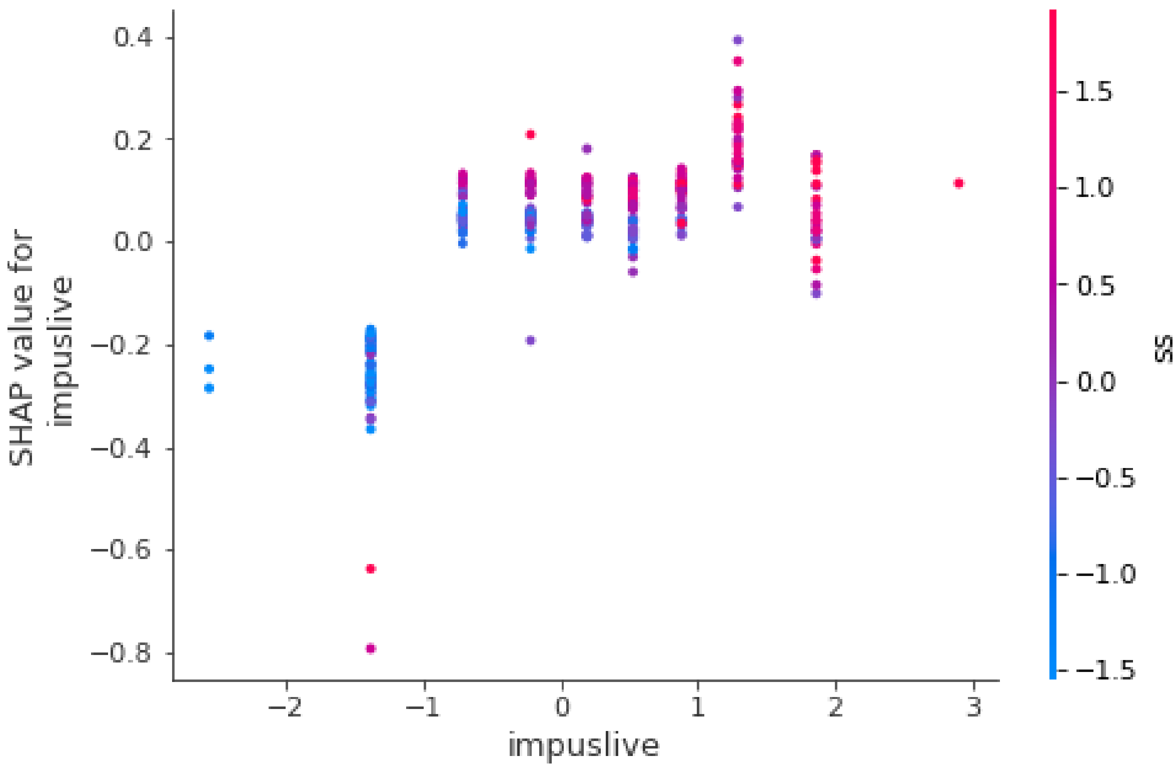

SHAP value for oscore
oscore
country



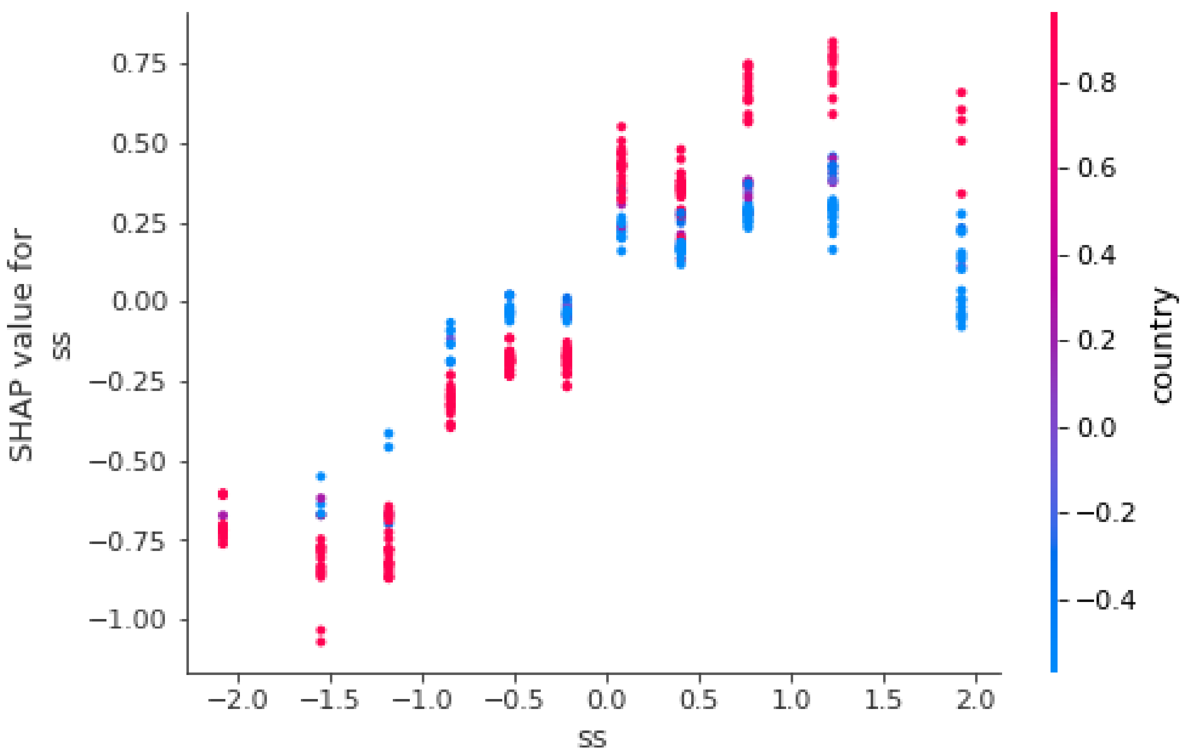
SHAP value for ascore
ascore
country



SHAP value for cscore
cscore
escore

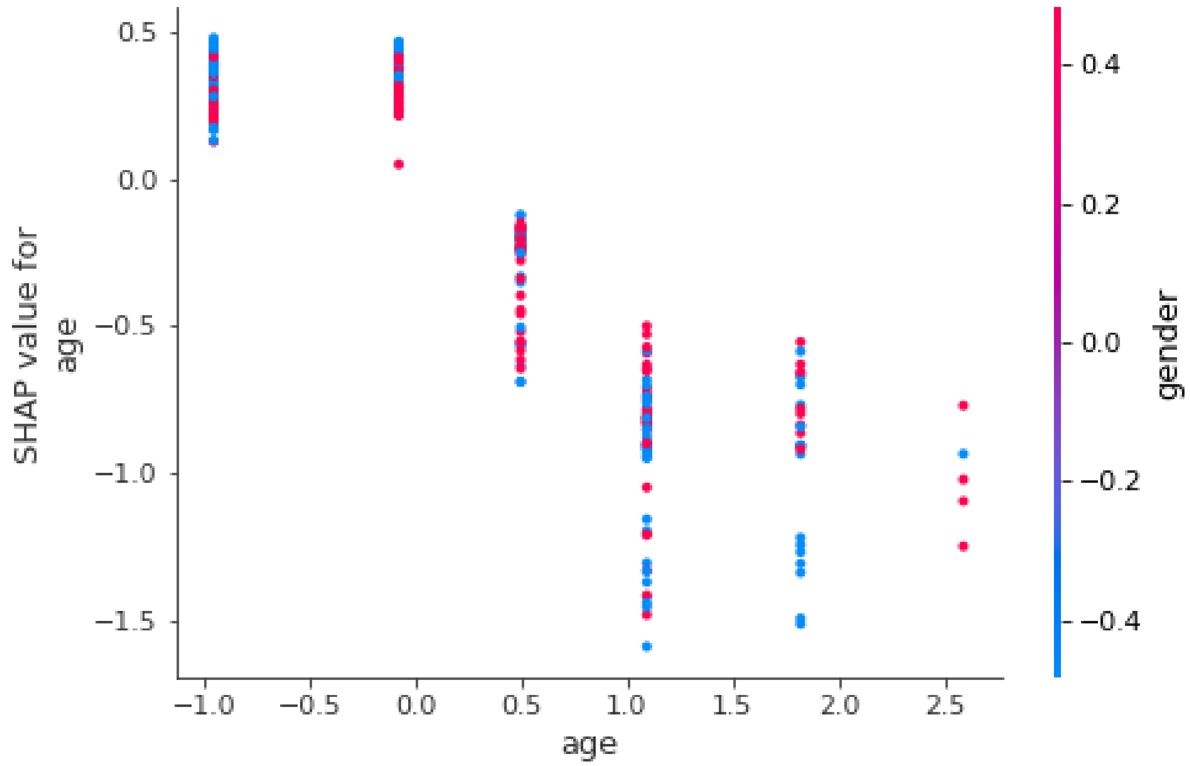


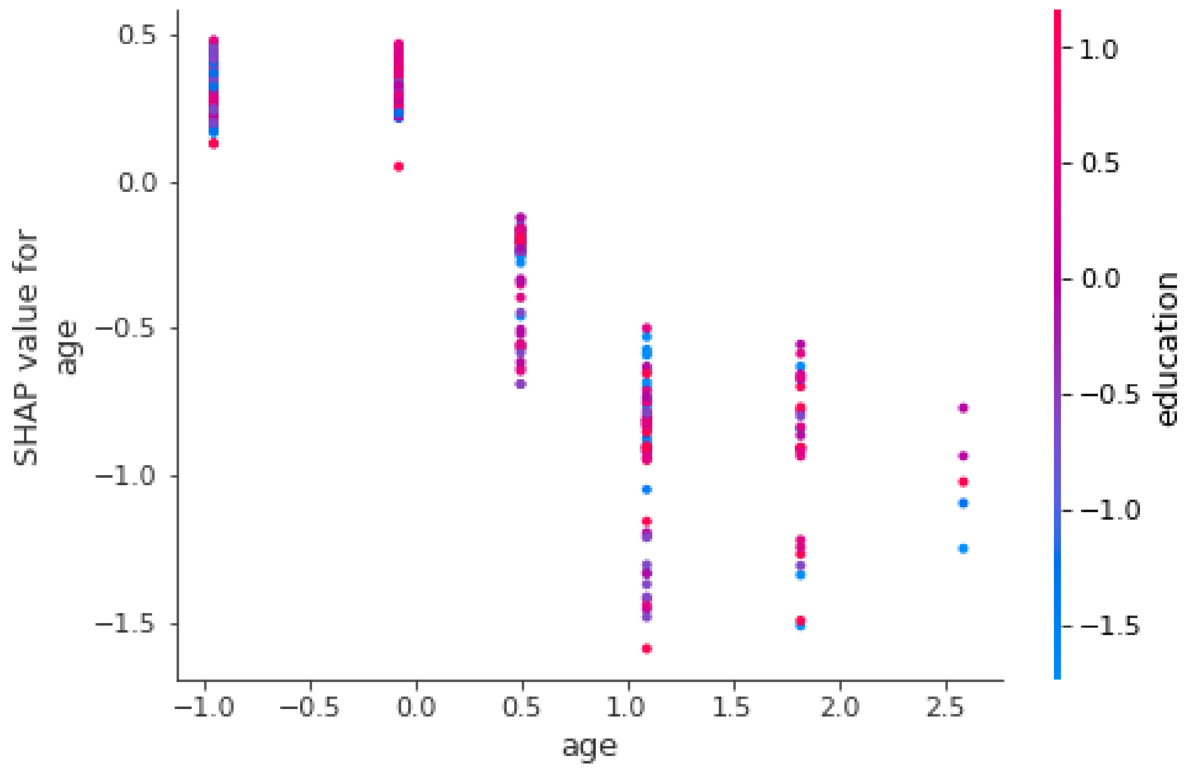
SHAP value for impuslive
impuslive
ss

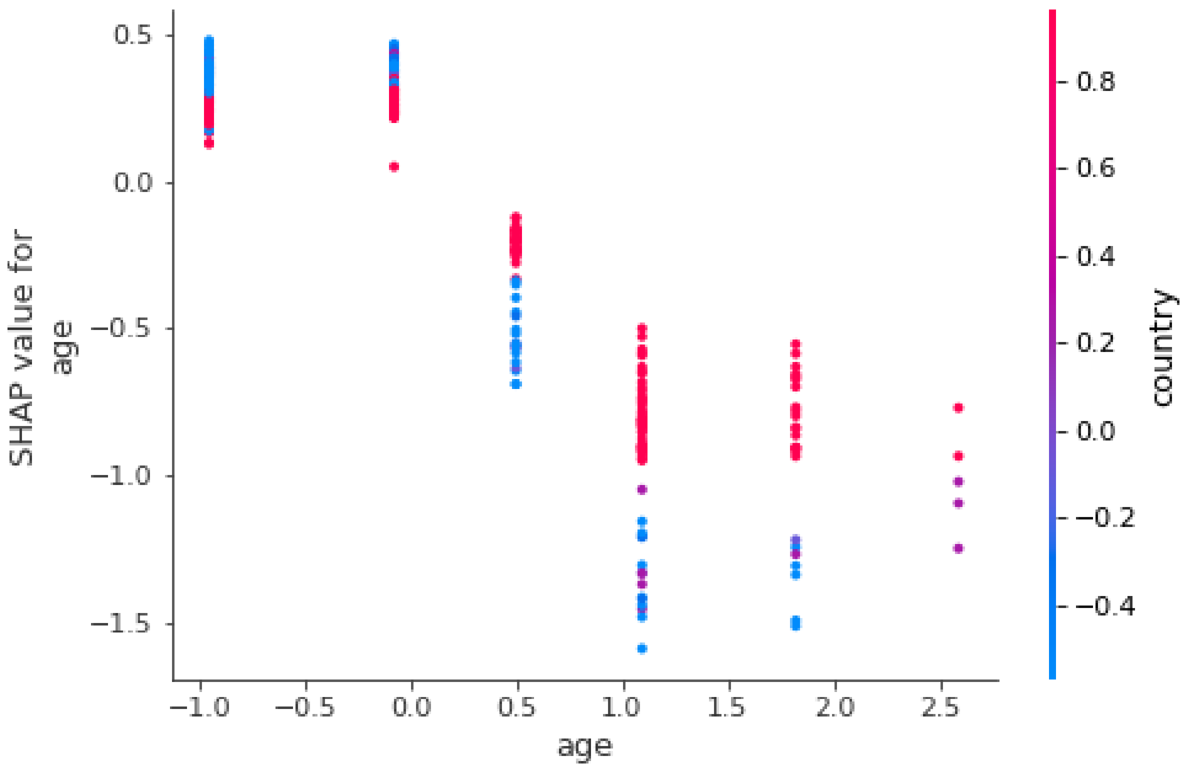


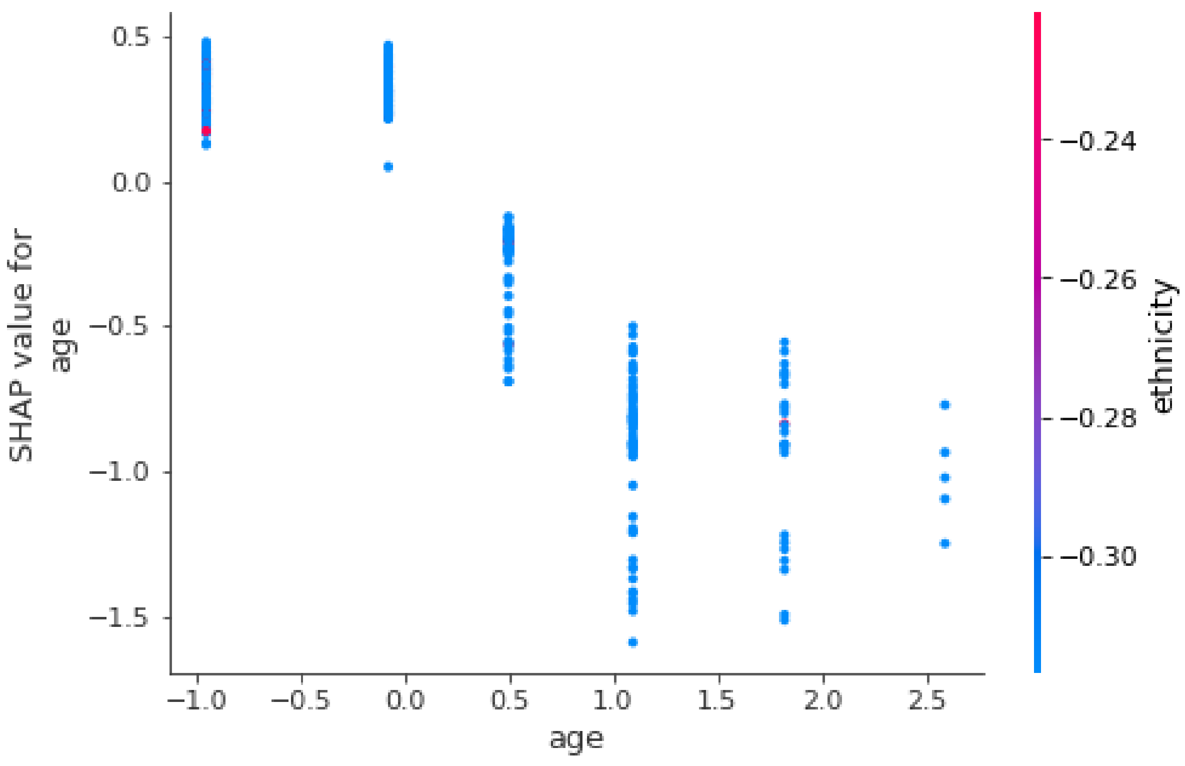
SHAP value for ss
ss
country

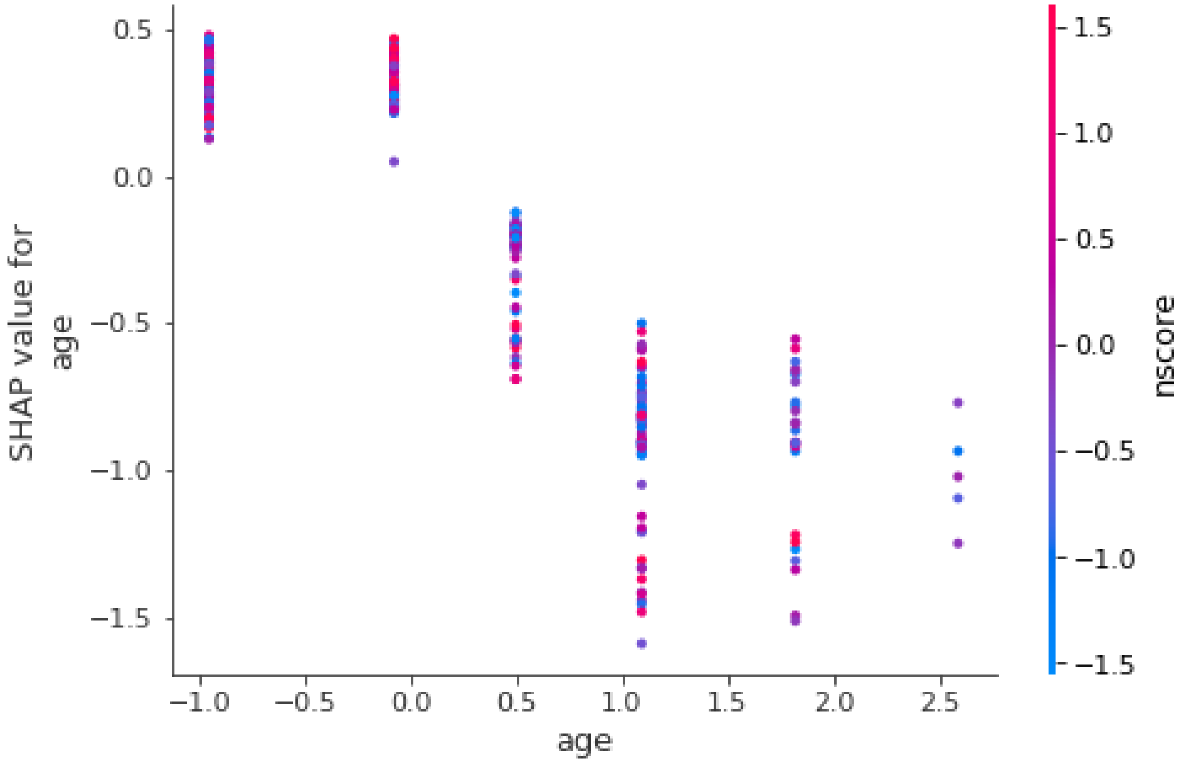
SHAP INTERSECTION Plot on Model GradientBoost with Dataset Mushroom Tes
t Set

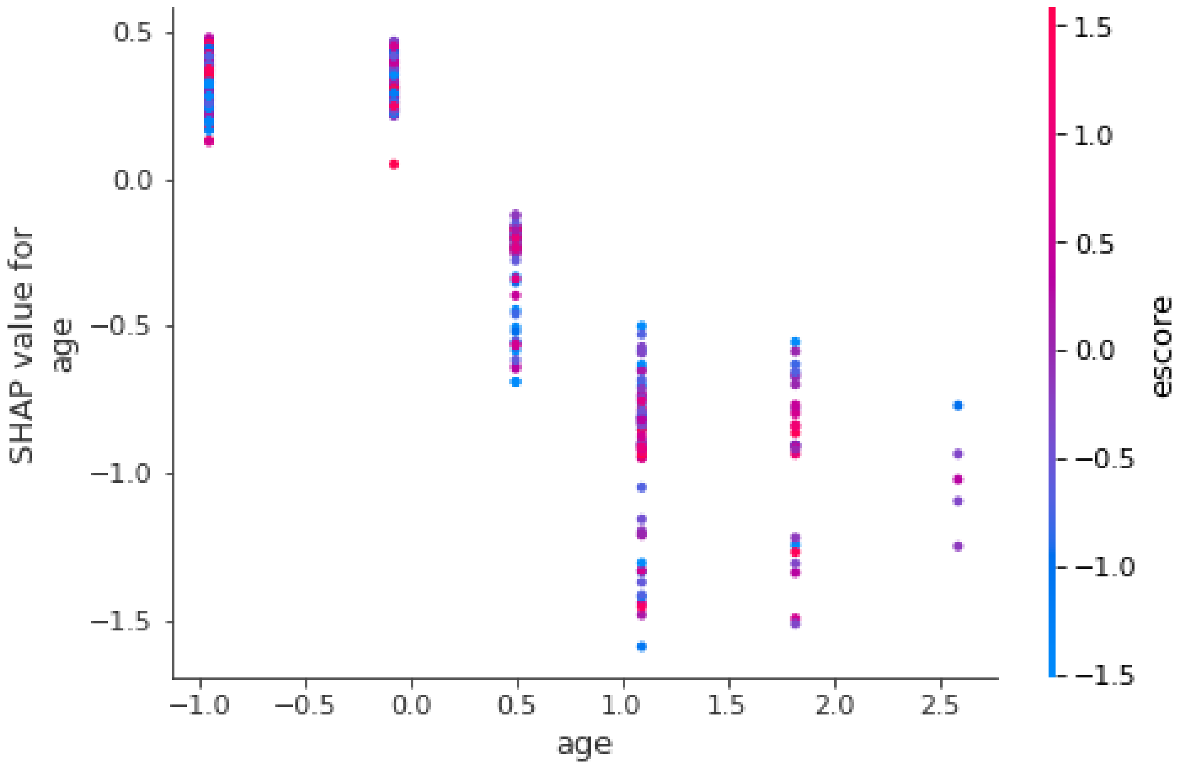


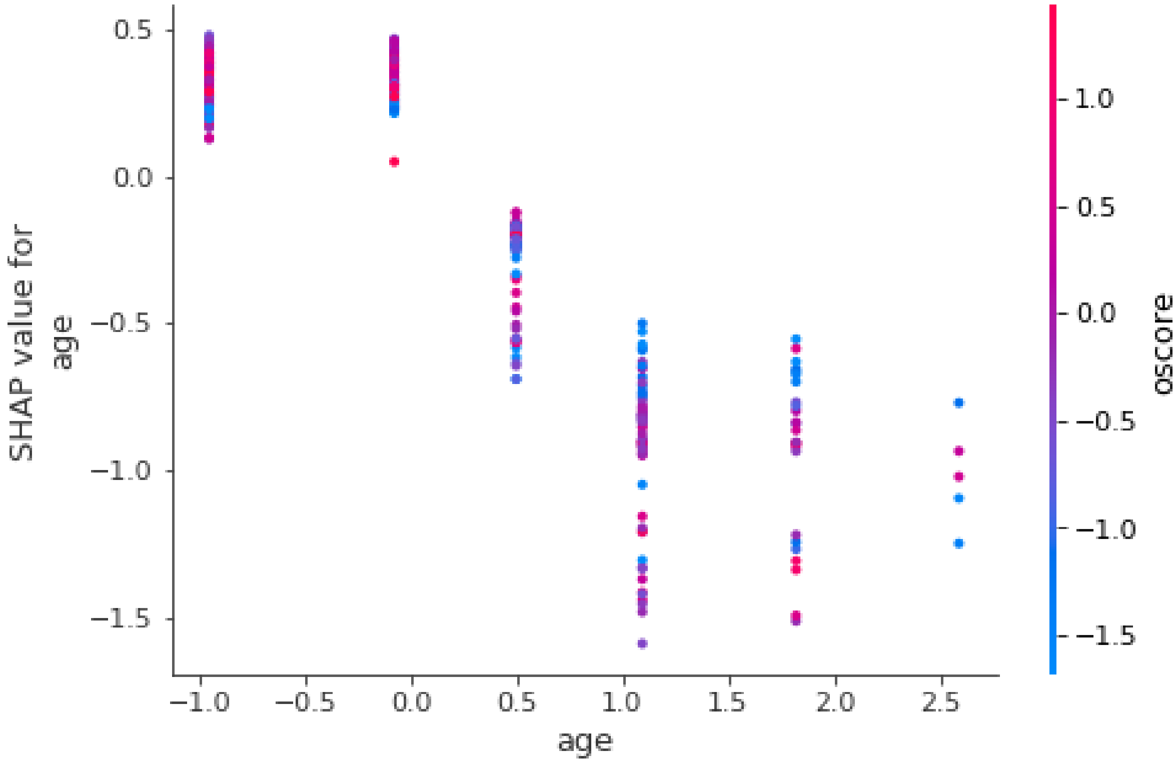


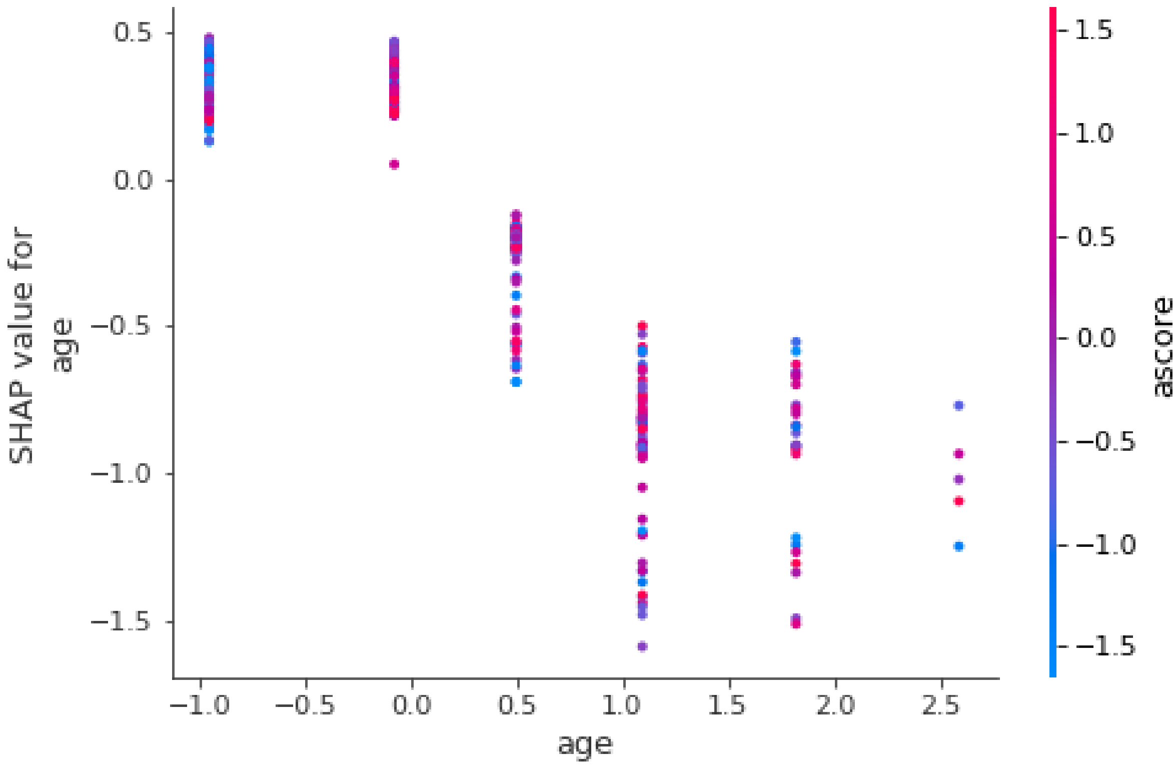

SHAP value for age
0.5
0.0
−0.5
−1.0
−1.5
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
age
country
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

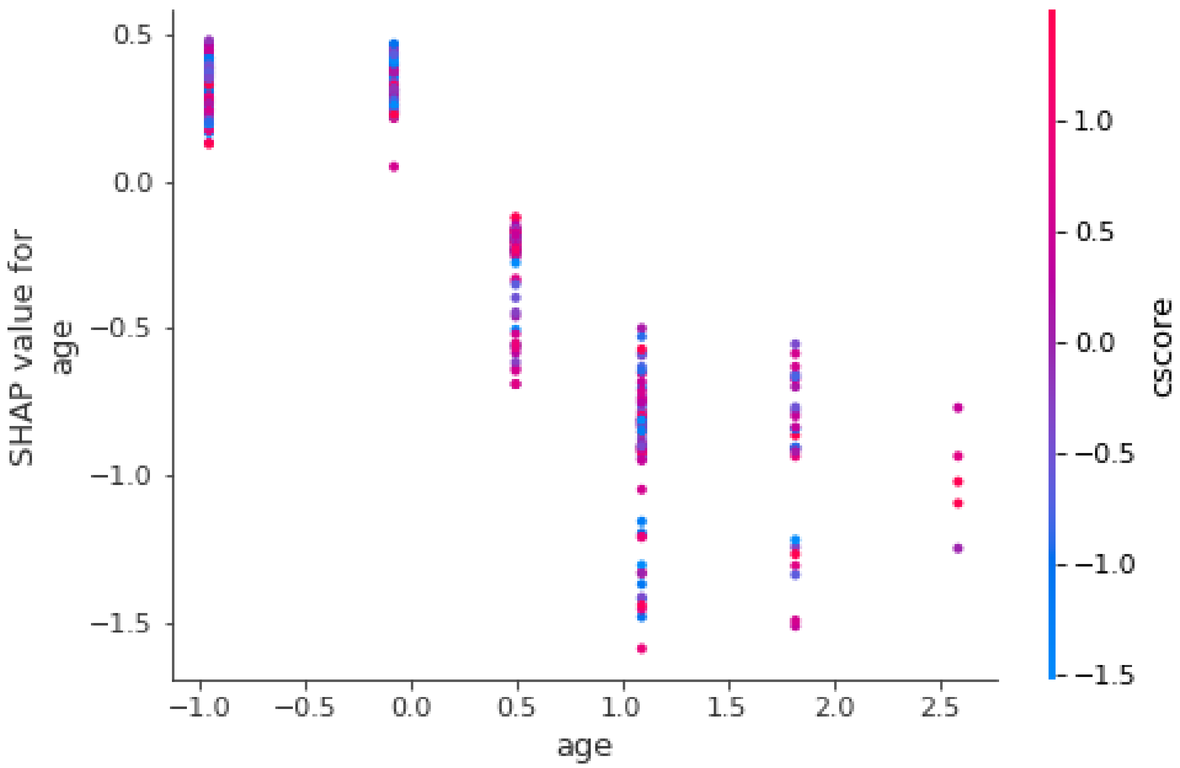

SHAP value for age
0.5
0.0
−0.5
−1.0
−1.5
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
age
ethnicity
−0.24
−0.26
−0.28
−0.30

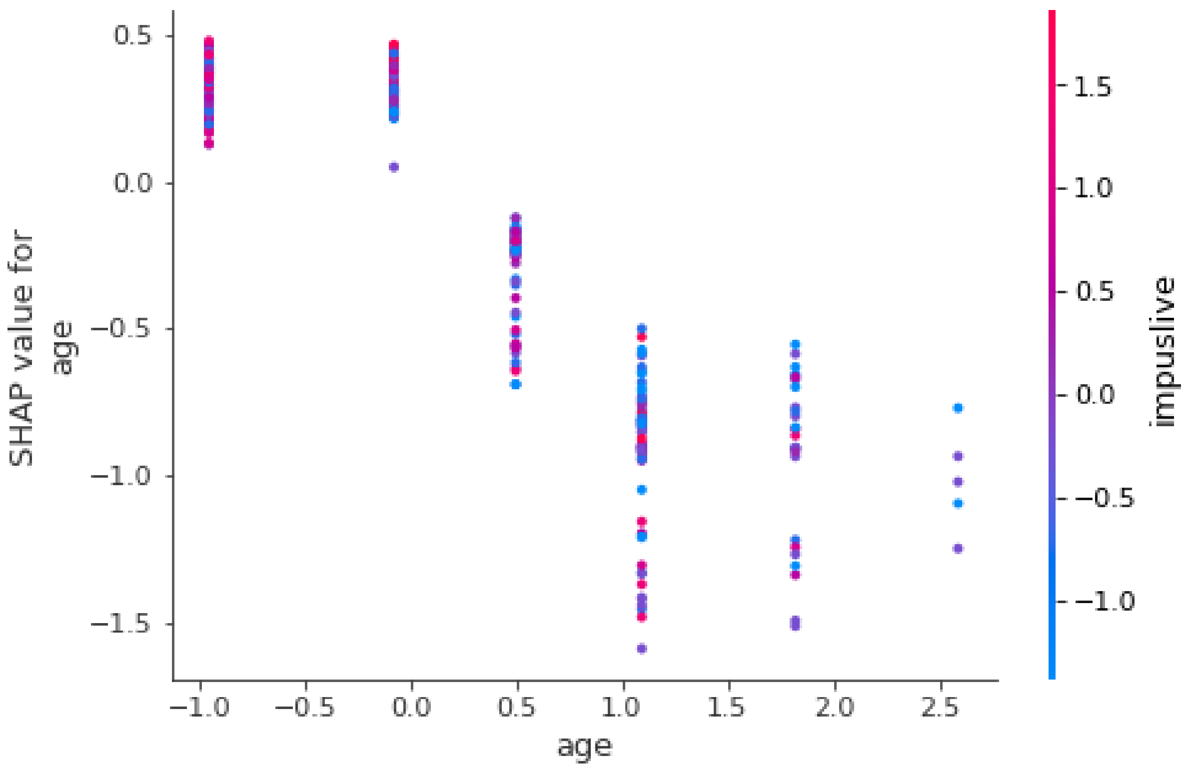

SHAP value for age
age
nscore
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

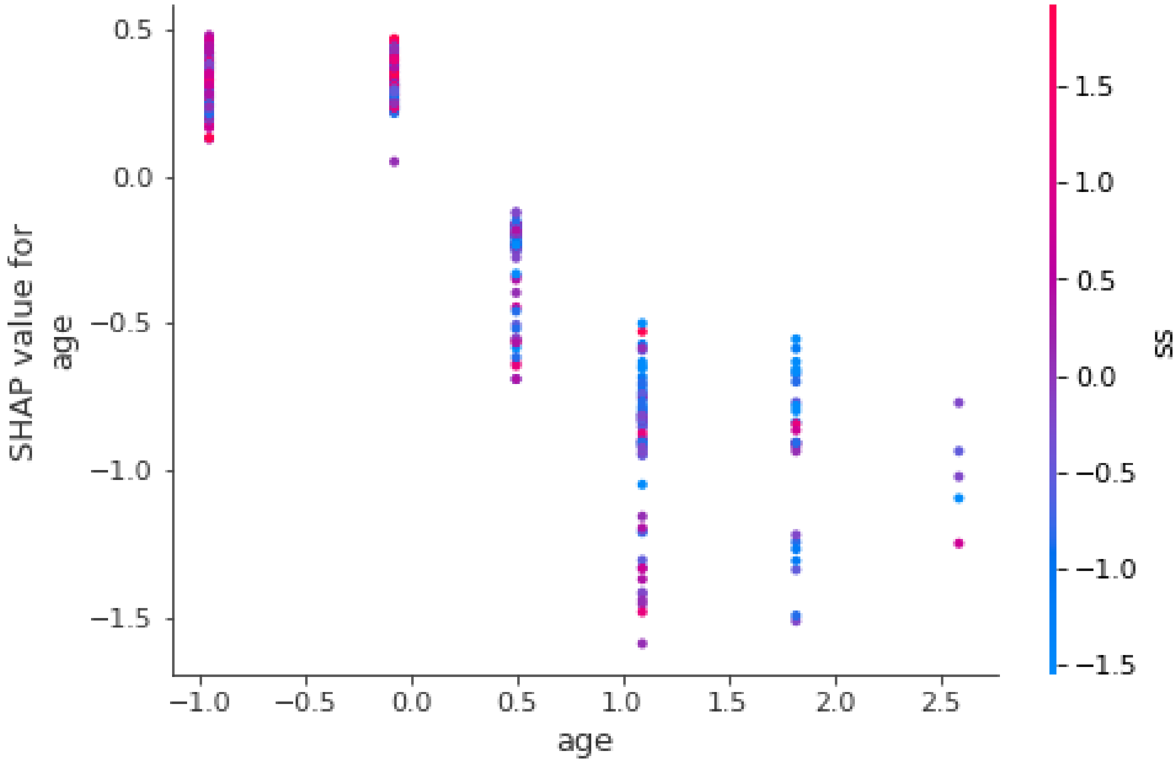

SHAP value for age
age
escore
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


SHAP value for age
age
oscore
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

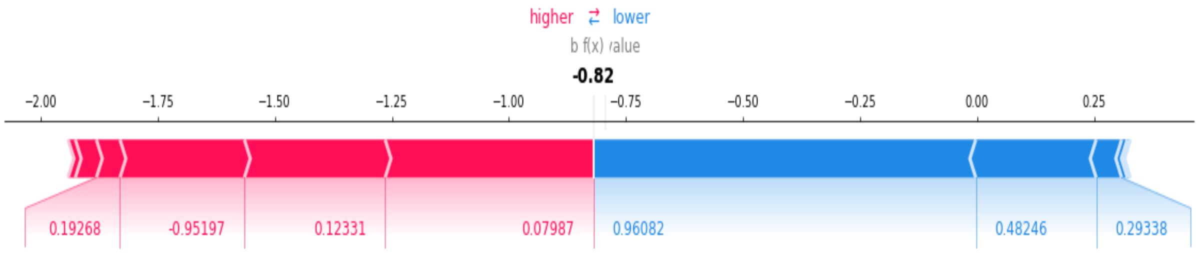

SHAP value for age
age
ascore
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


SHAP value for age
age
cscore

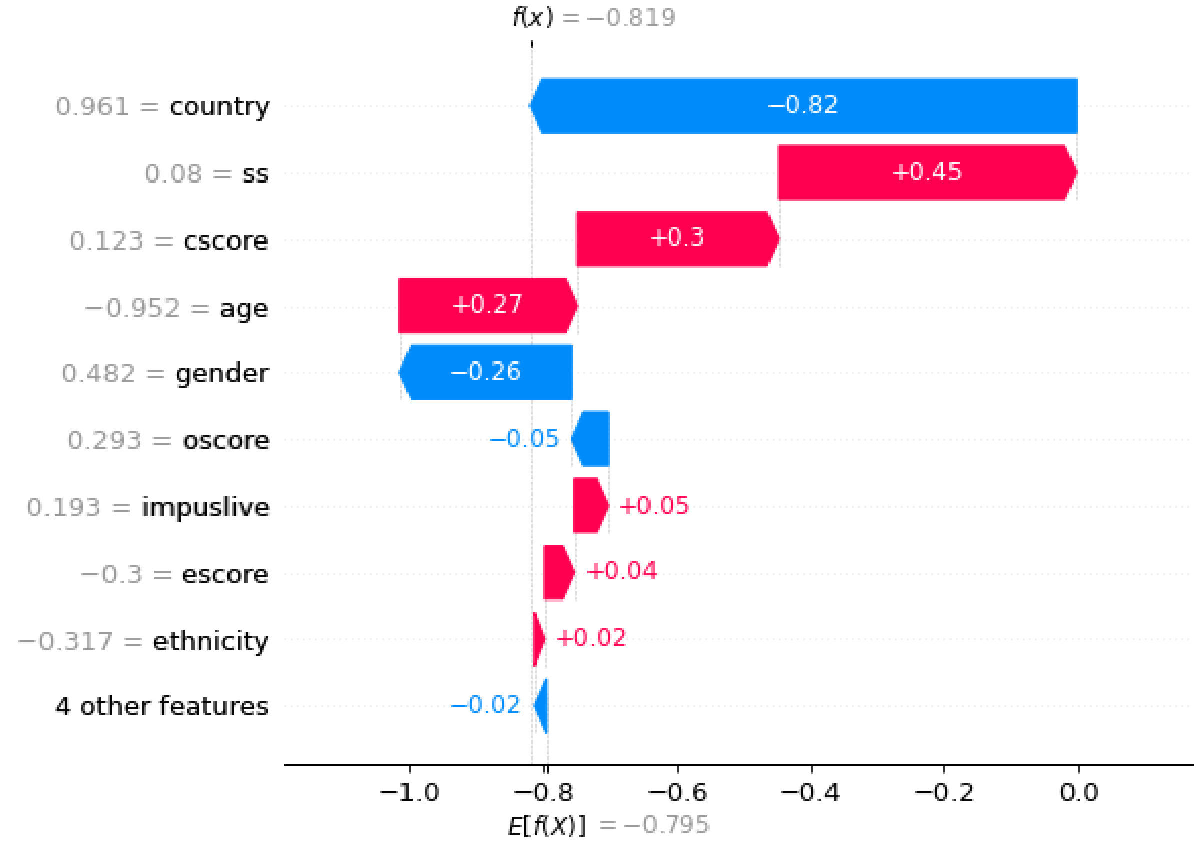

SHAP value for age
age
impuslive



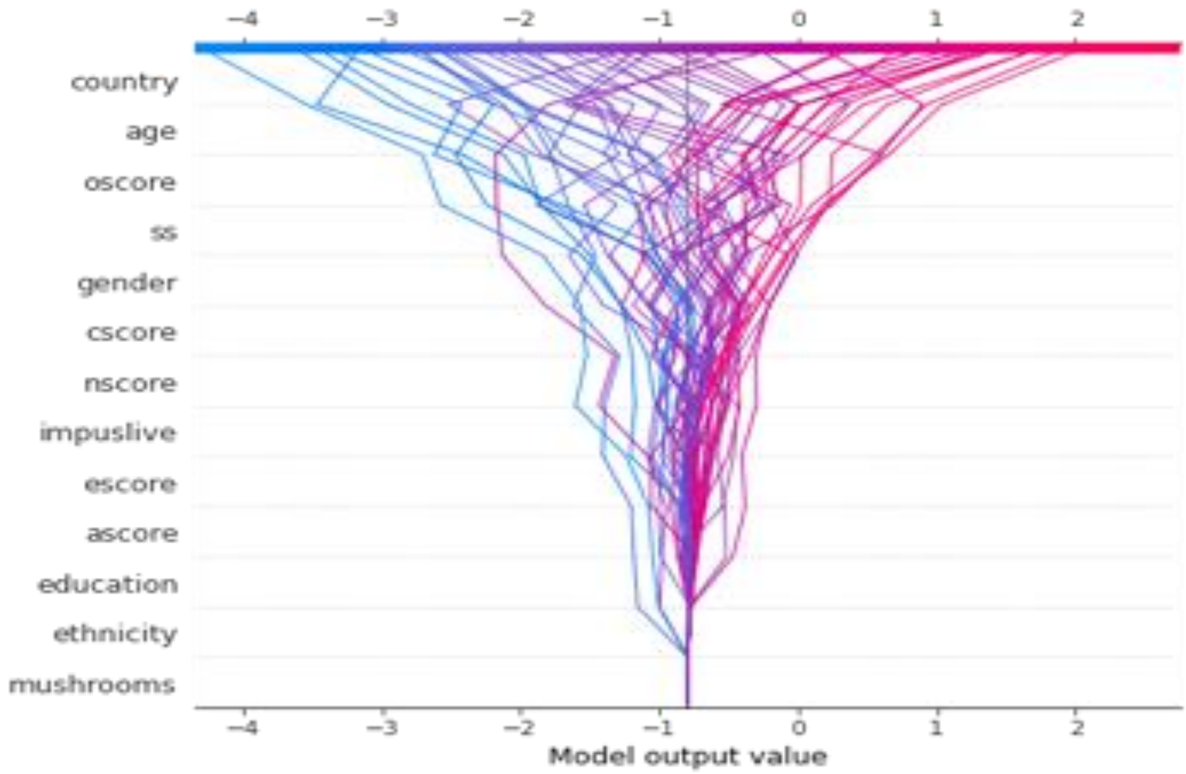
SHAP Force Plot for a single prediction in Model GradientBoost



SHAP Waterfall Plot for a single prediction in Model GradientBoost

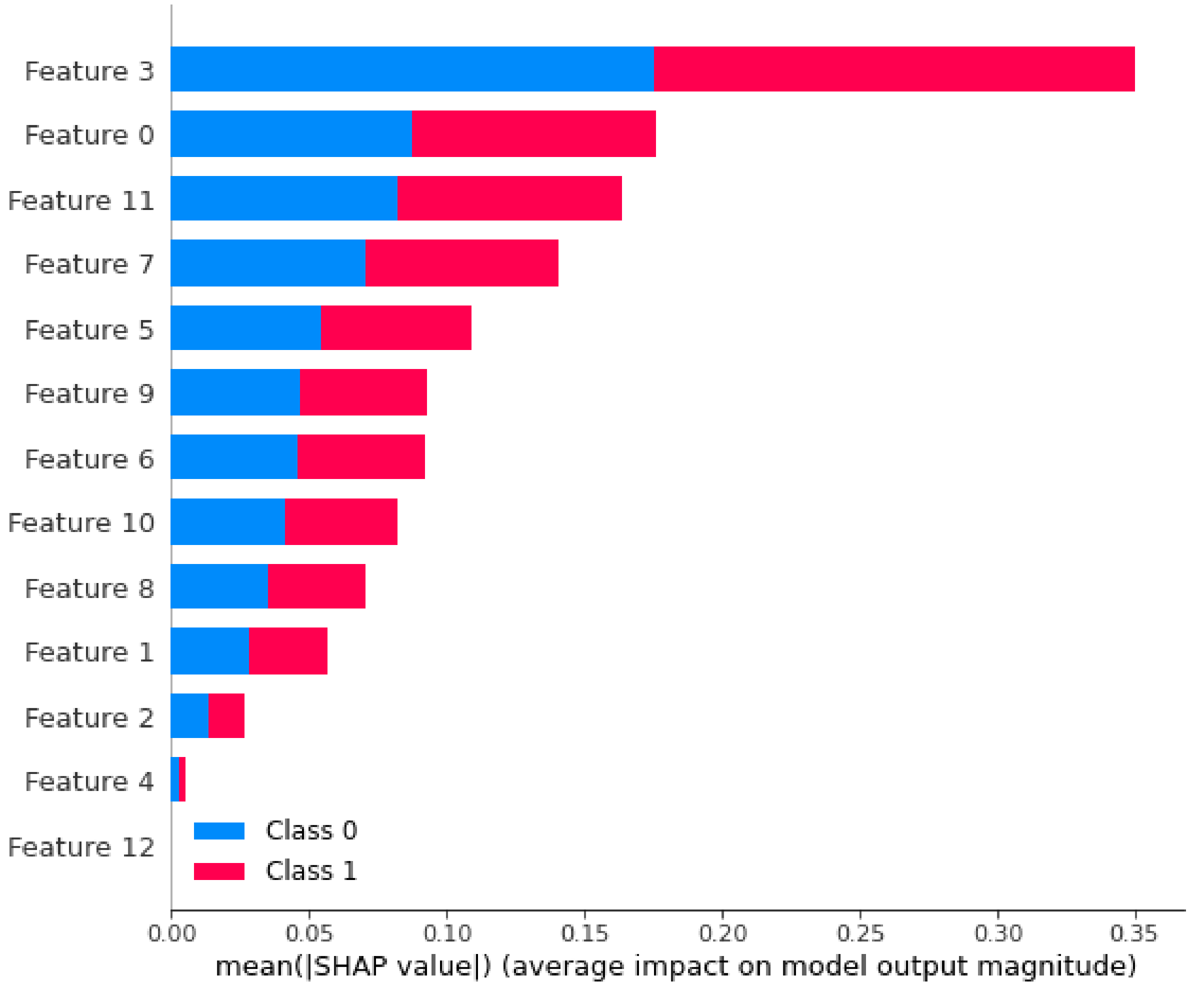


SHAP Decision Plot for Model GradientBoost on Dataset Mushroom Test Set

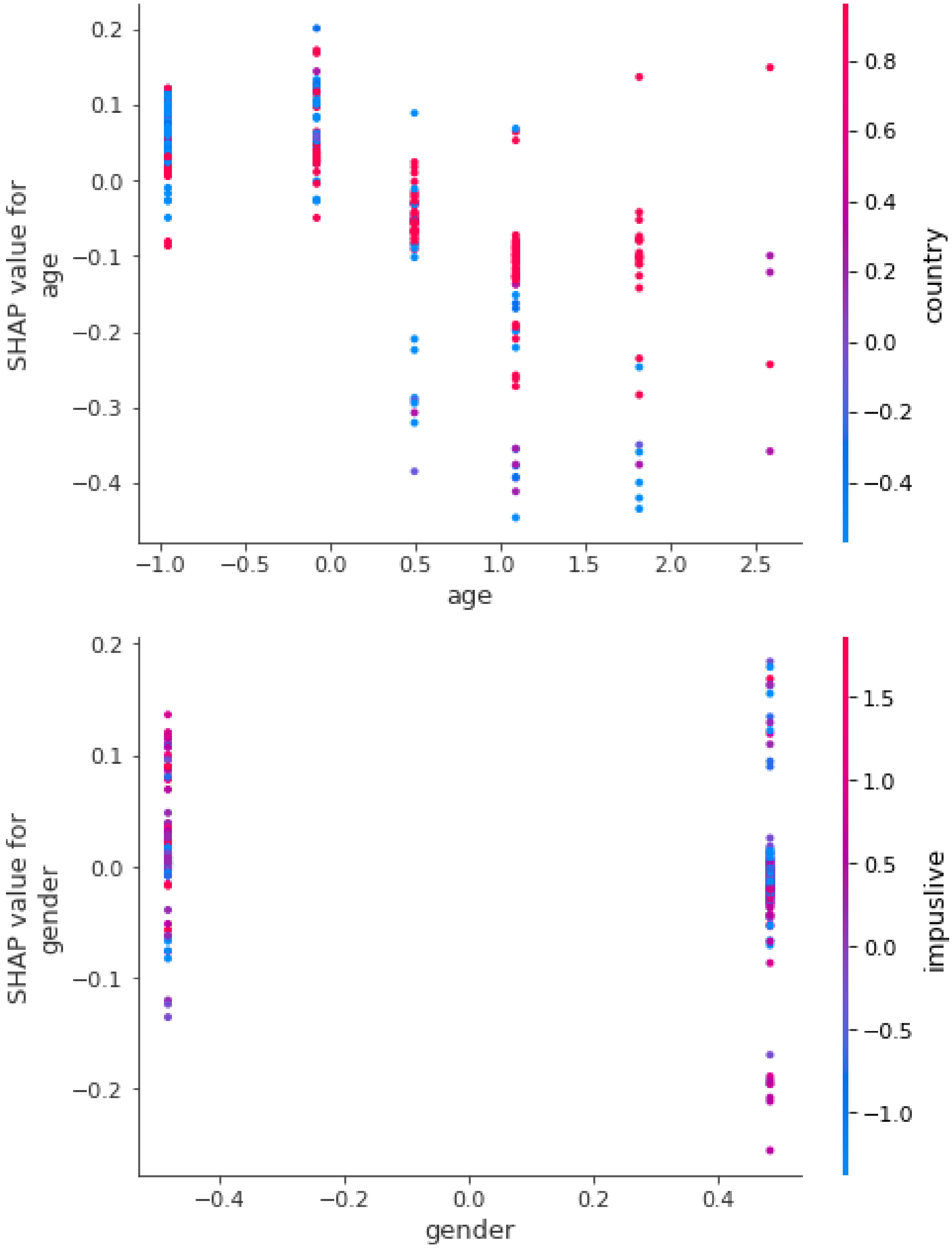


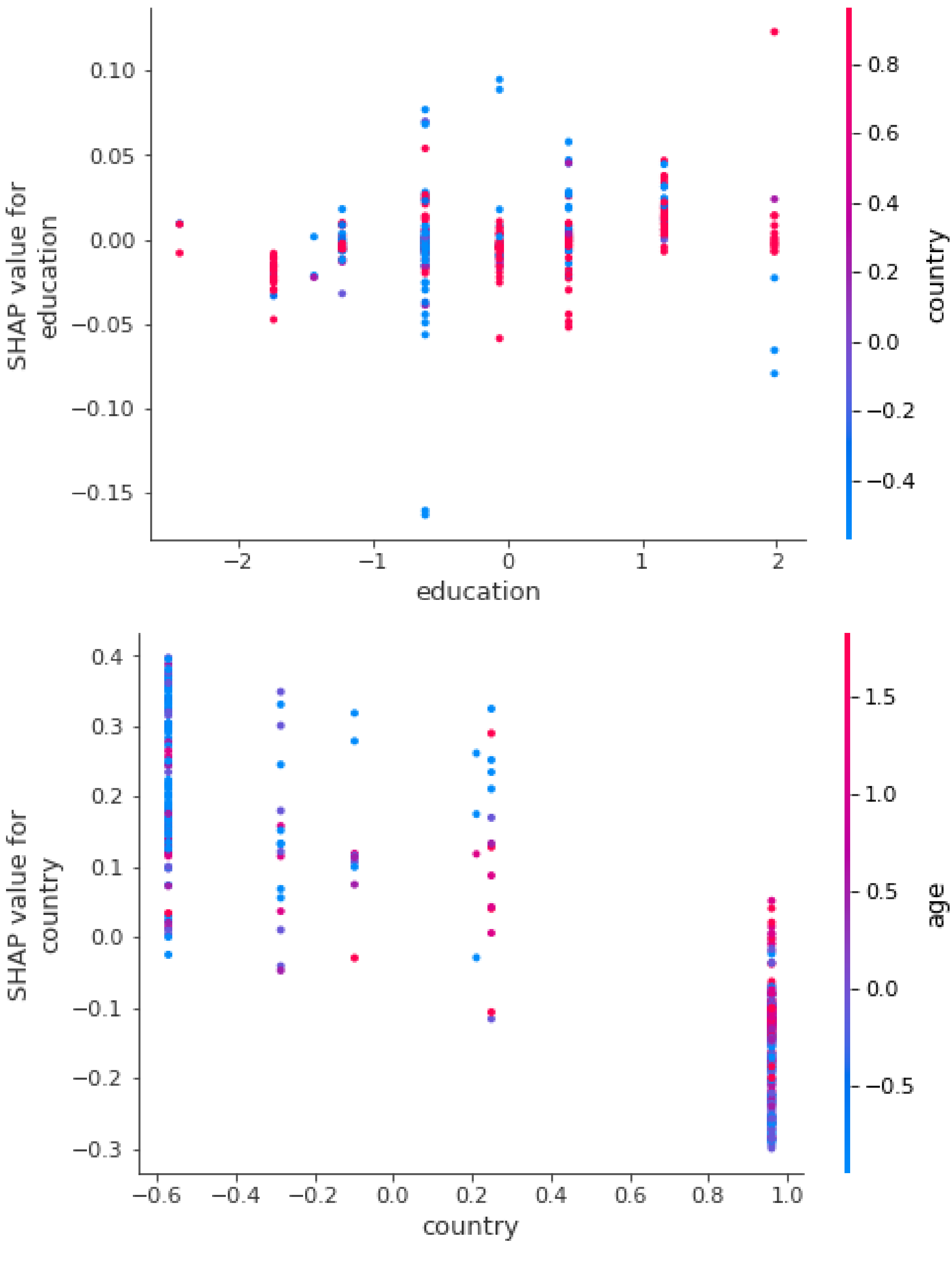
Computing SHAP values with the worst model (Decision Tree) on MUSHROOM test set

SHAP Summary Plot for Model Decision Tree on Dataset Mushroom Test Set

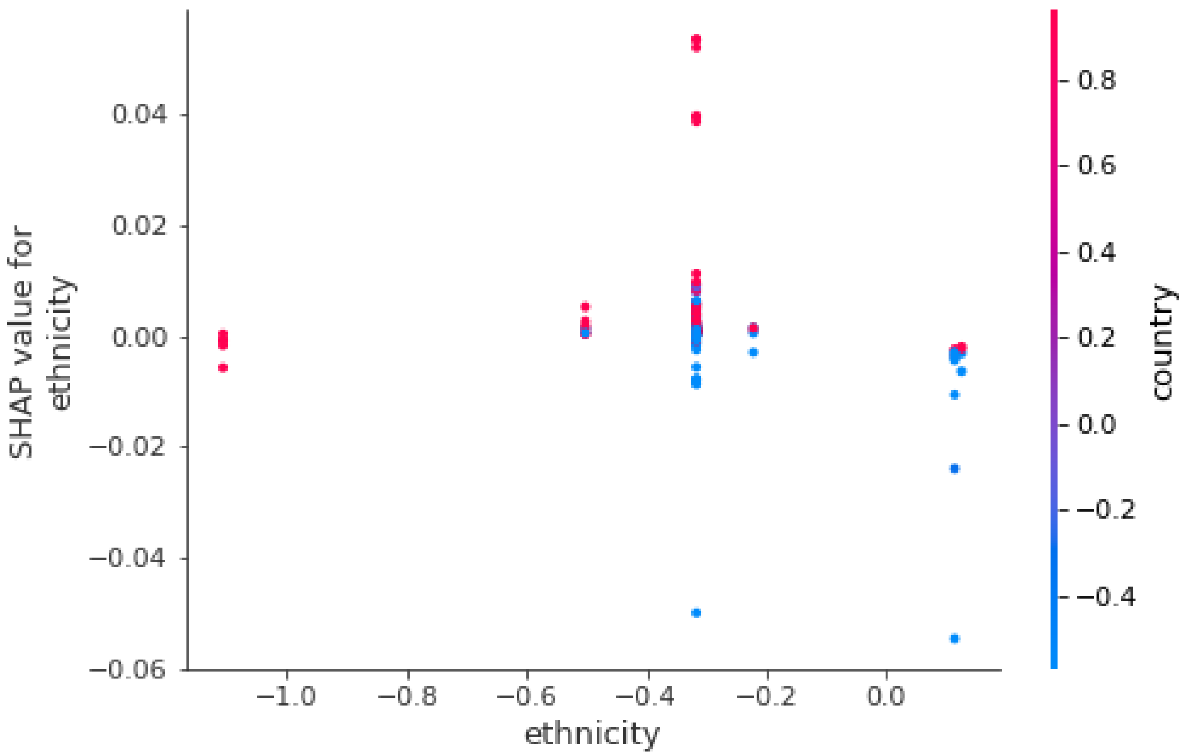


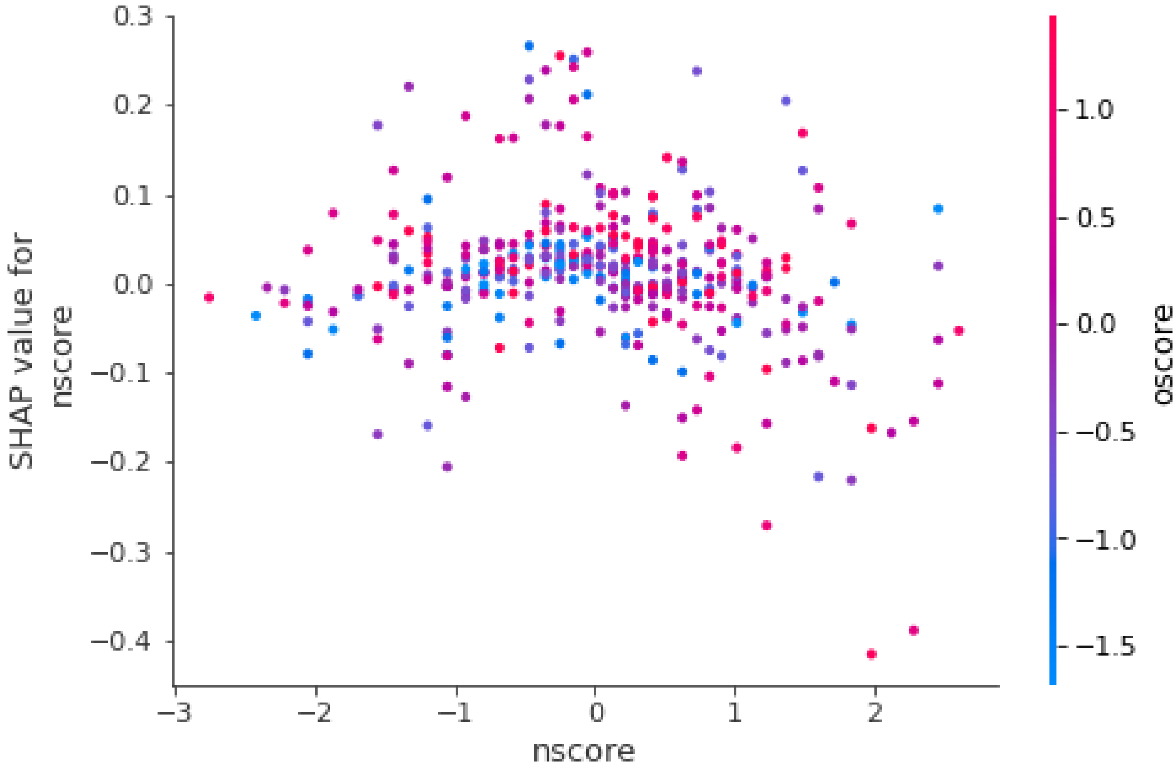
SHAP Dependence Plot on Model Decision Tree with Dataset Mushroom Test Set

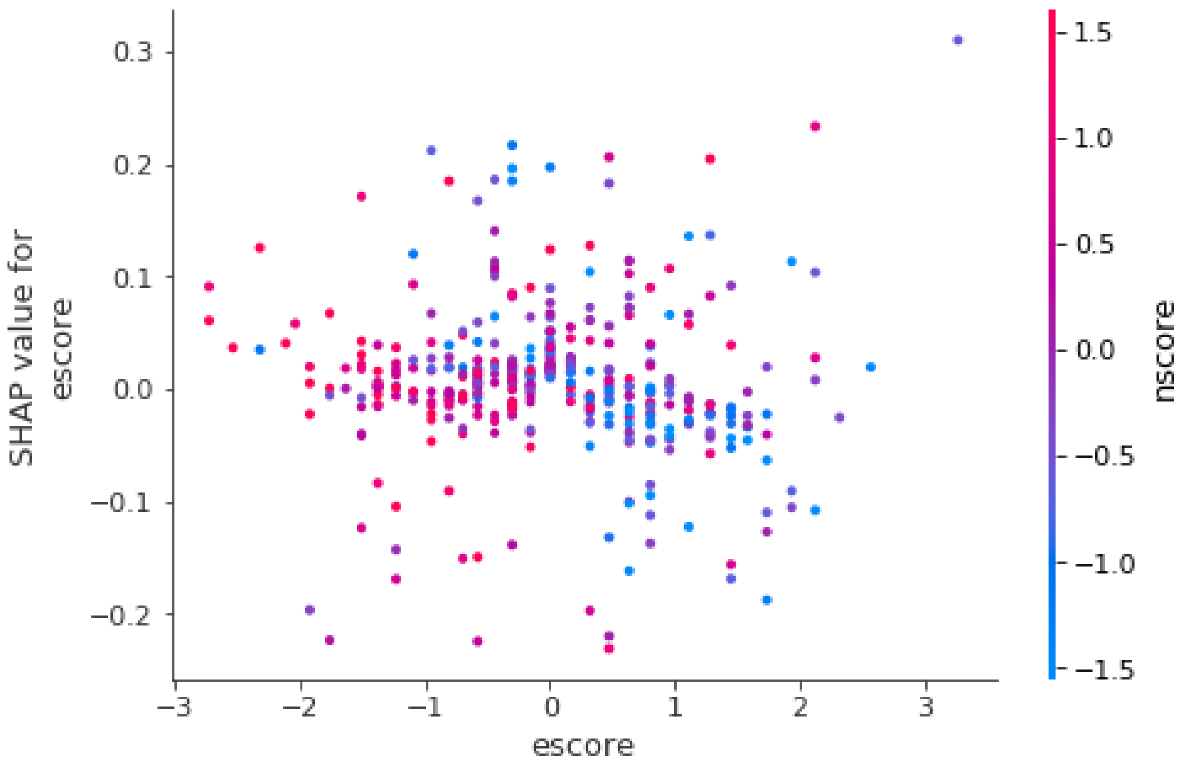


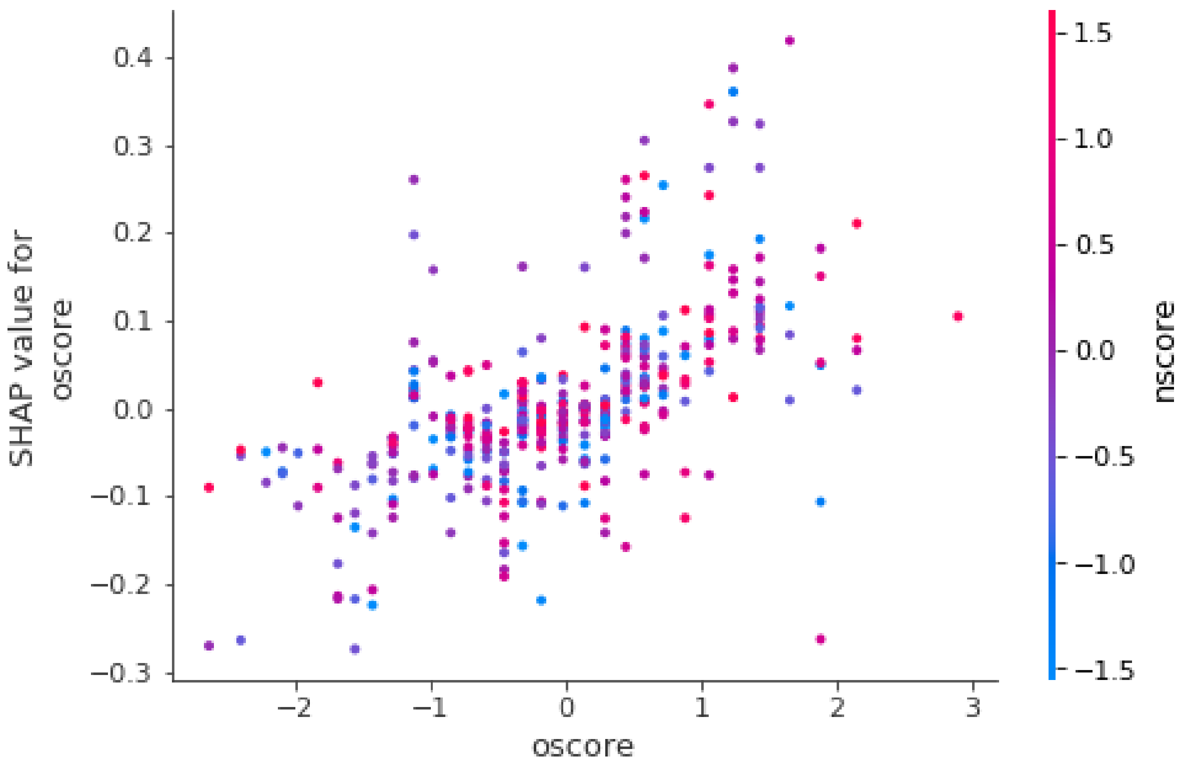


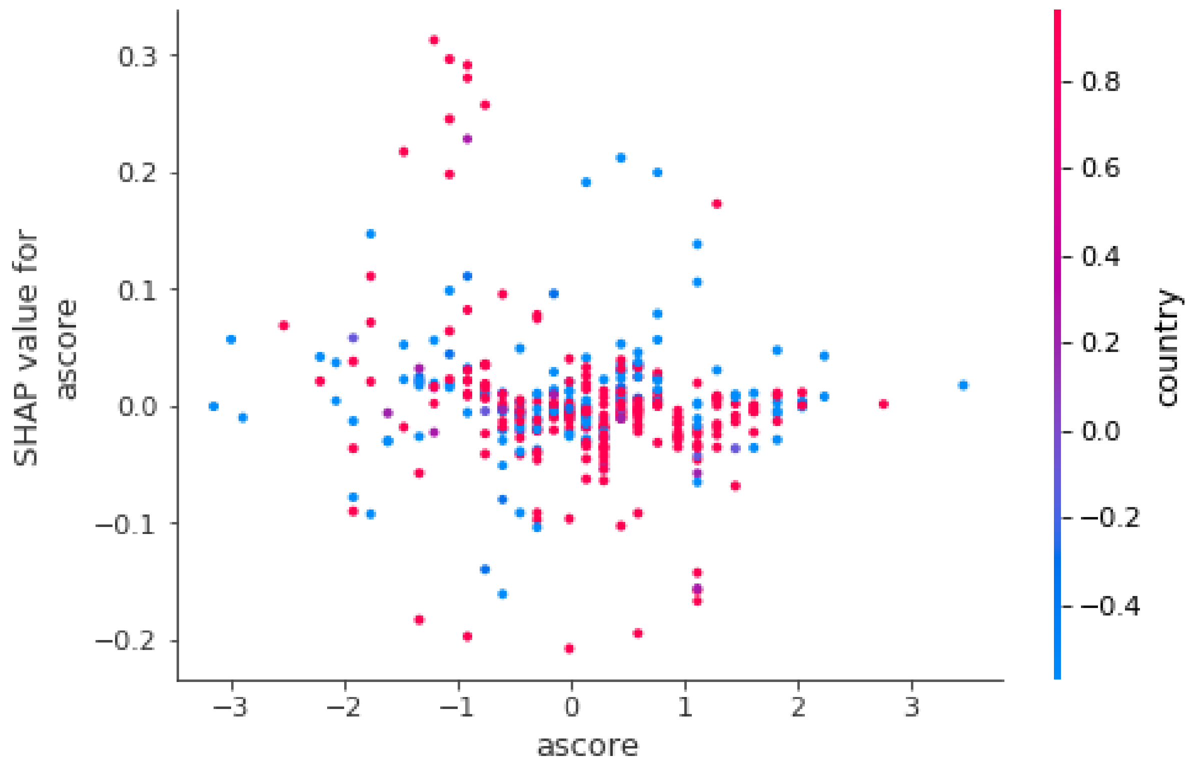
SHAP value for education
education
country
SHAP value for country
country
age

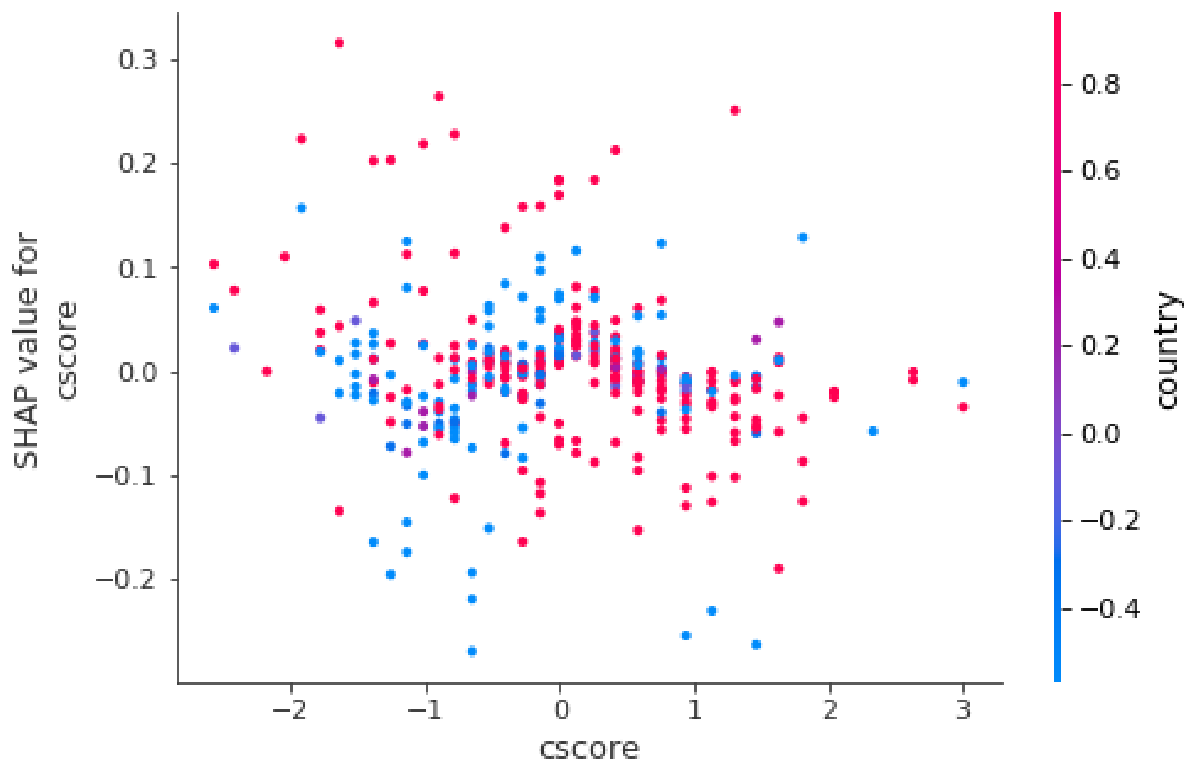

SHAP value for ethnicity
ethnicity
country
0.04
0.02
0.00
−0.02
−0.04
−0.06
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

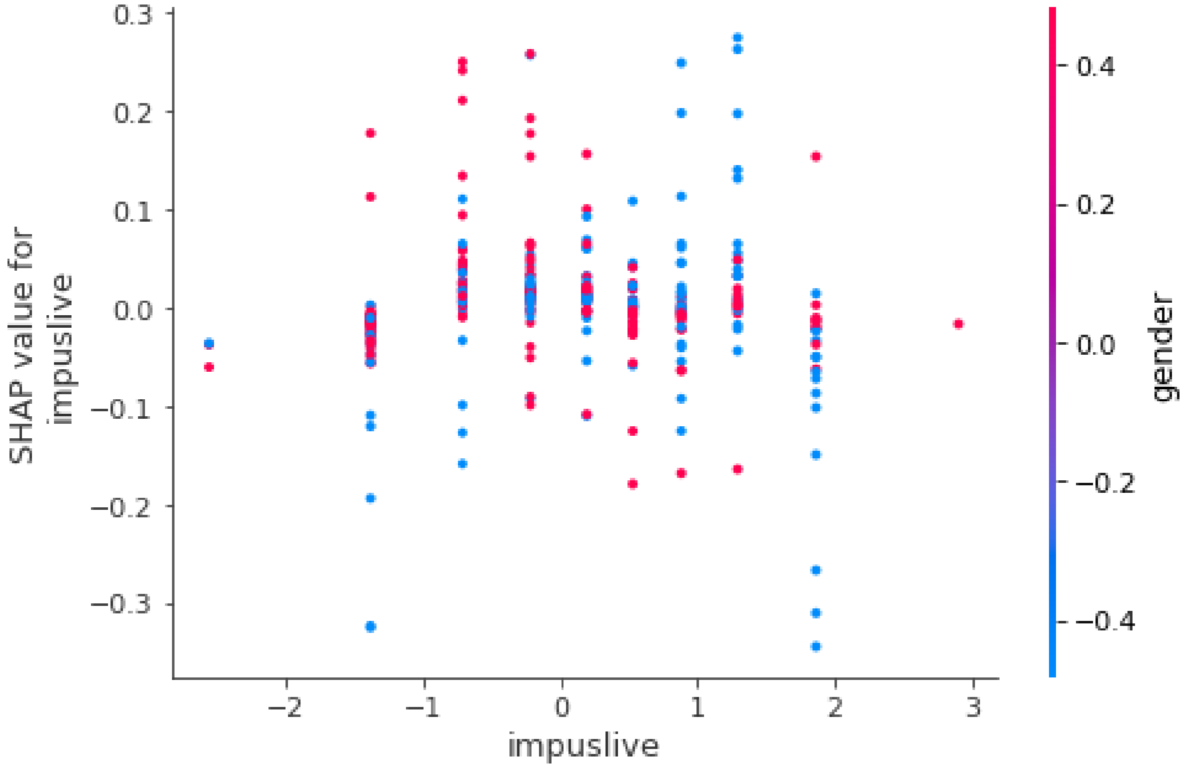

SHAP value for nscore
nscore
oscore
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−3
−2
−1
0
1
2
1.0
0.5
0.0
−0.5
−1.0
−1.5

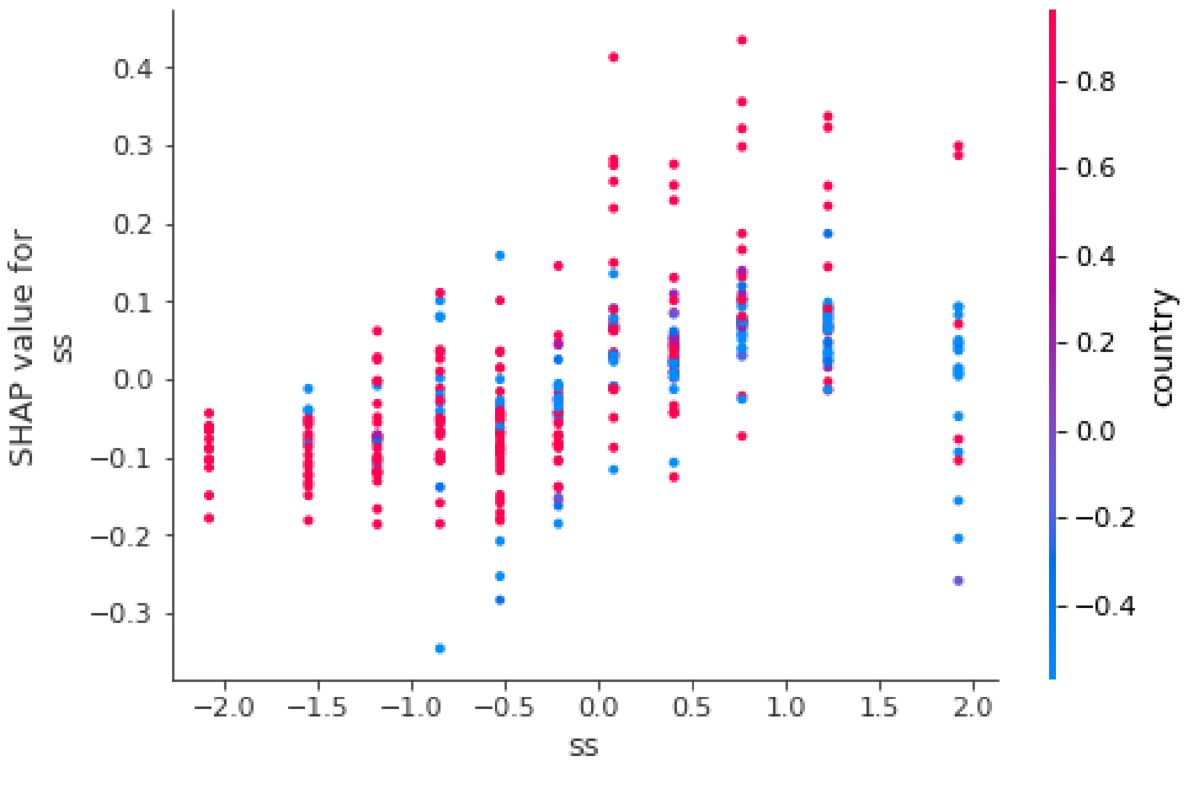

SHAP value for escore
escore
nscore
0.3
0.2
0.1
0.0
−0.1
−0.2
−3
−2
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


SHAP value for oscore
oscore
nscore
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−2
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

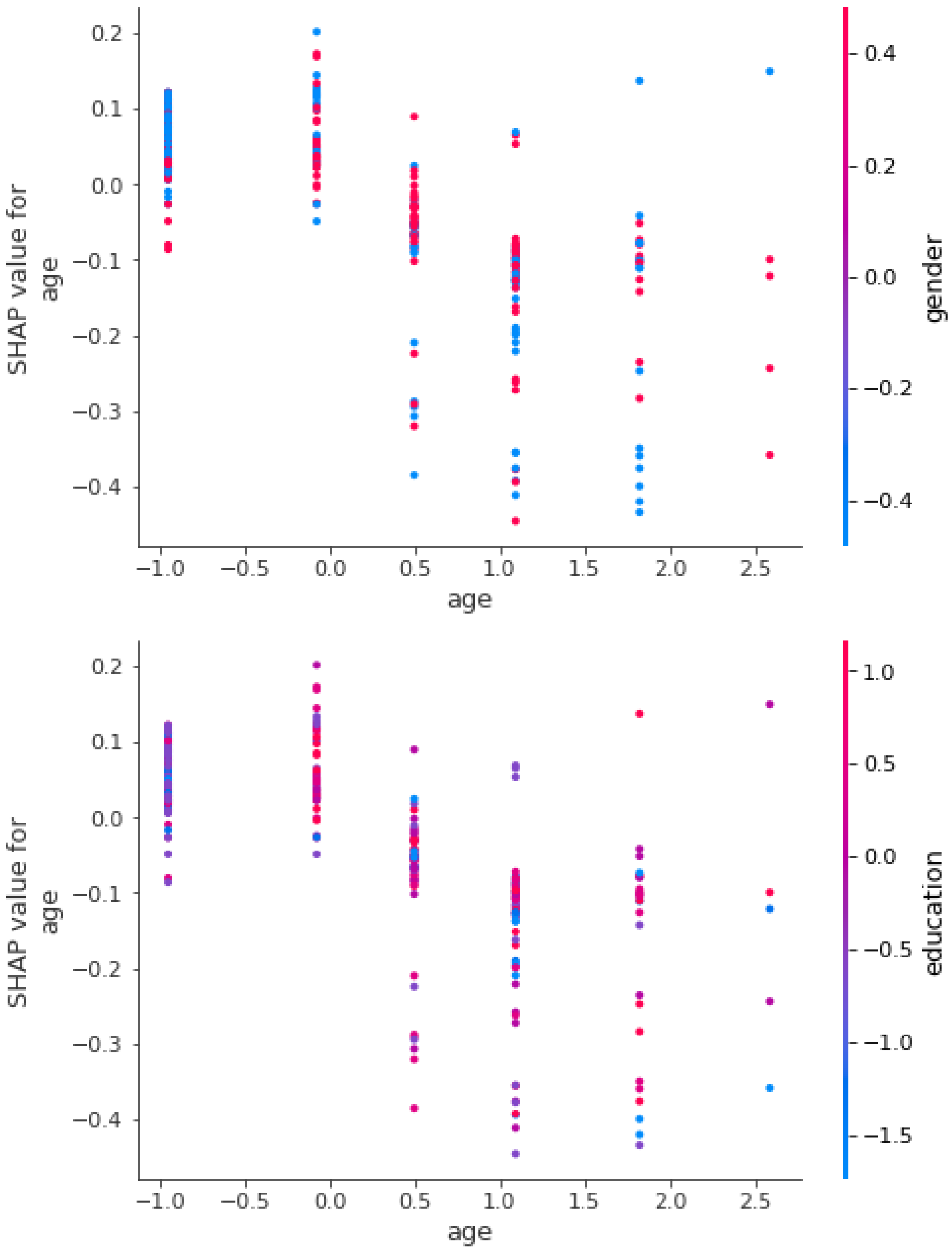

SHAP value for ascore
ascore
country

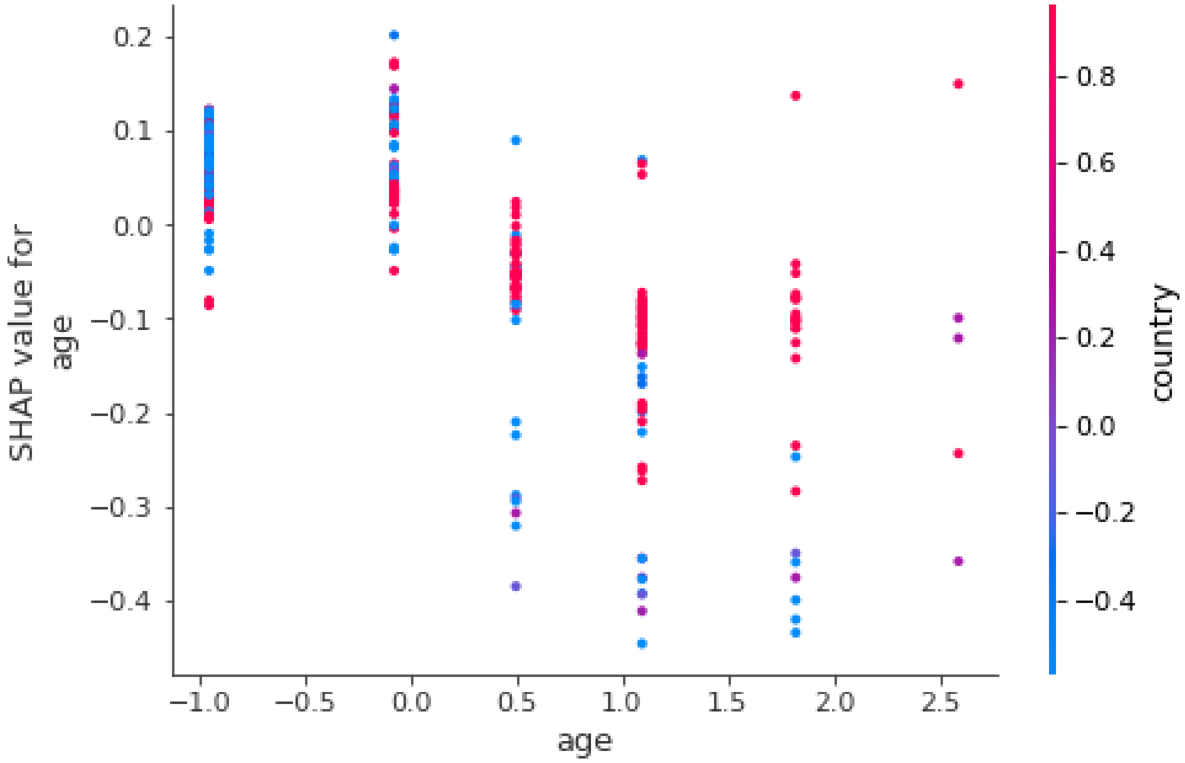

SHAP value for cscore
cscore
country

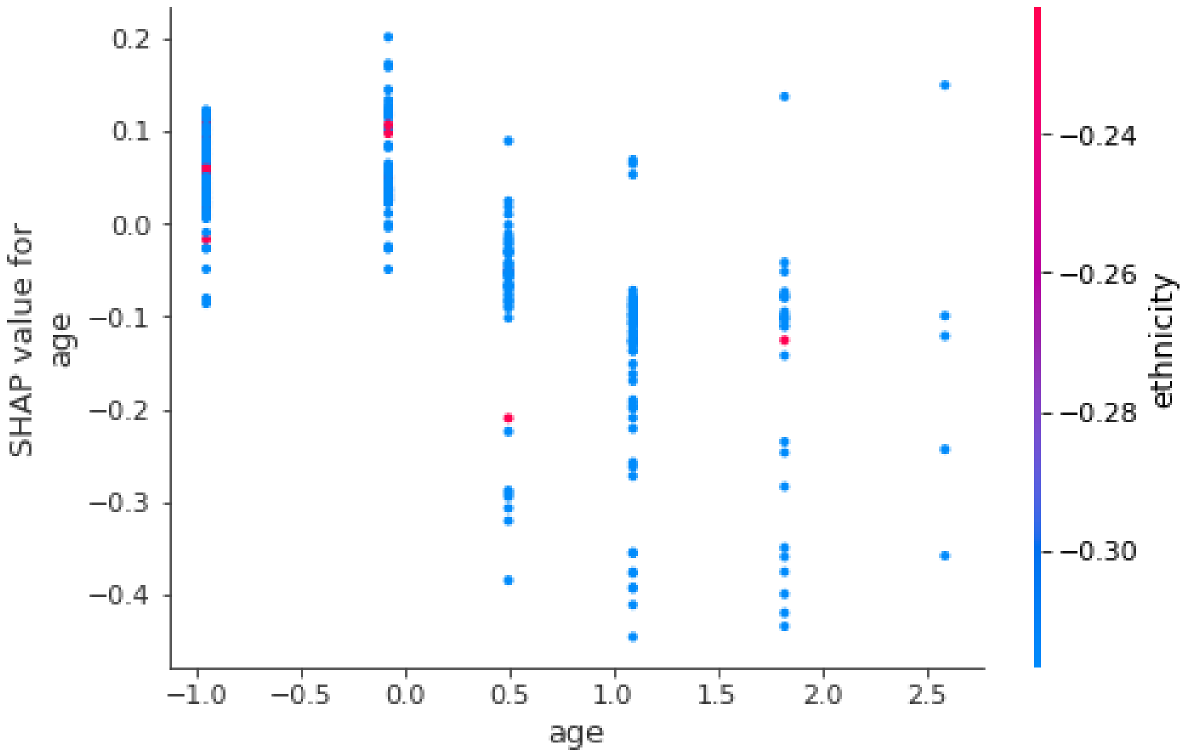

SHAP value for impuslive
impuslive
gender
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−2
−1
0
1
2
3
0.4
0.2
0.0
−0.2
−0.4

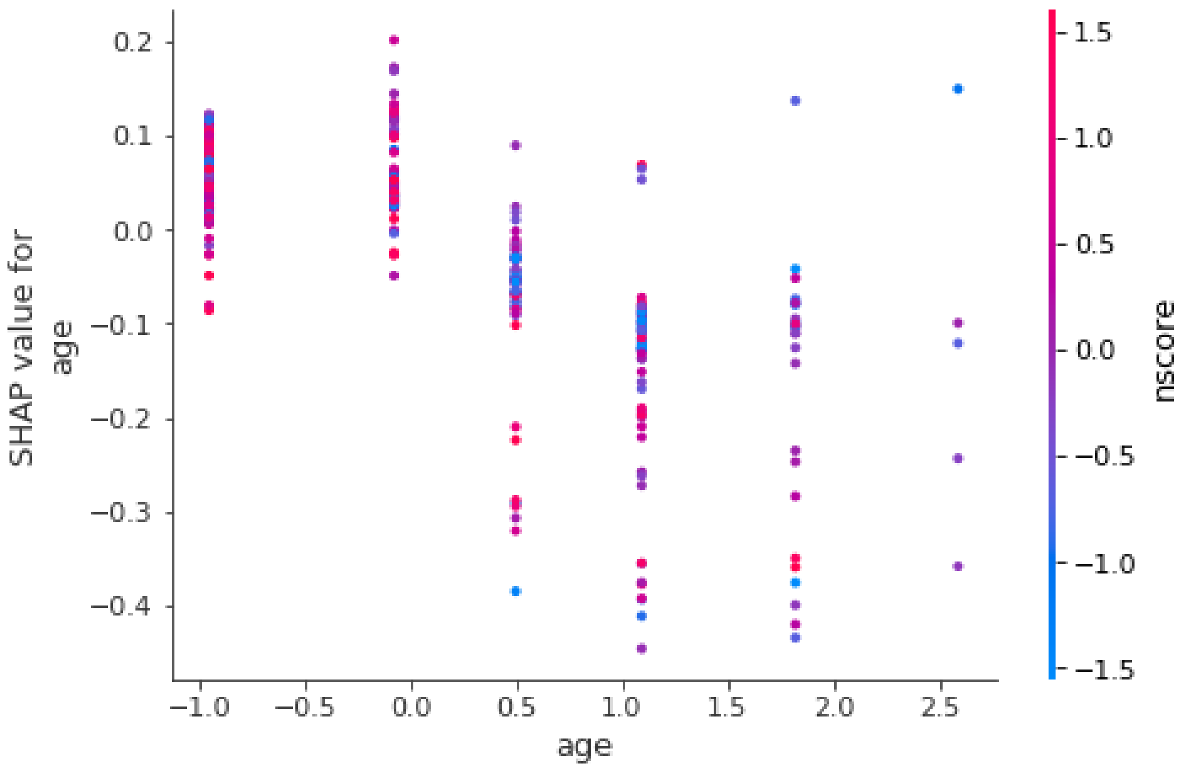

SHAP value for ss
ss
country
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

SHAP INTERSECTION Plot on Model Decision Tree with Dataset Mushroom Test Set

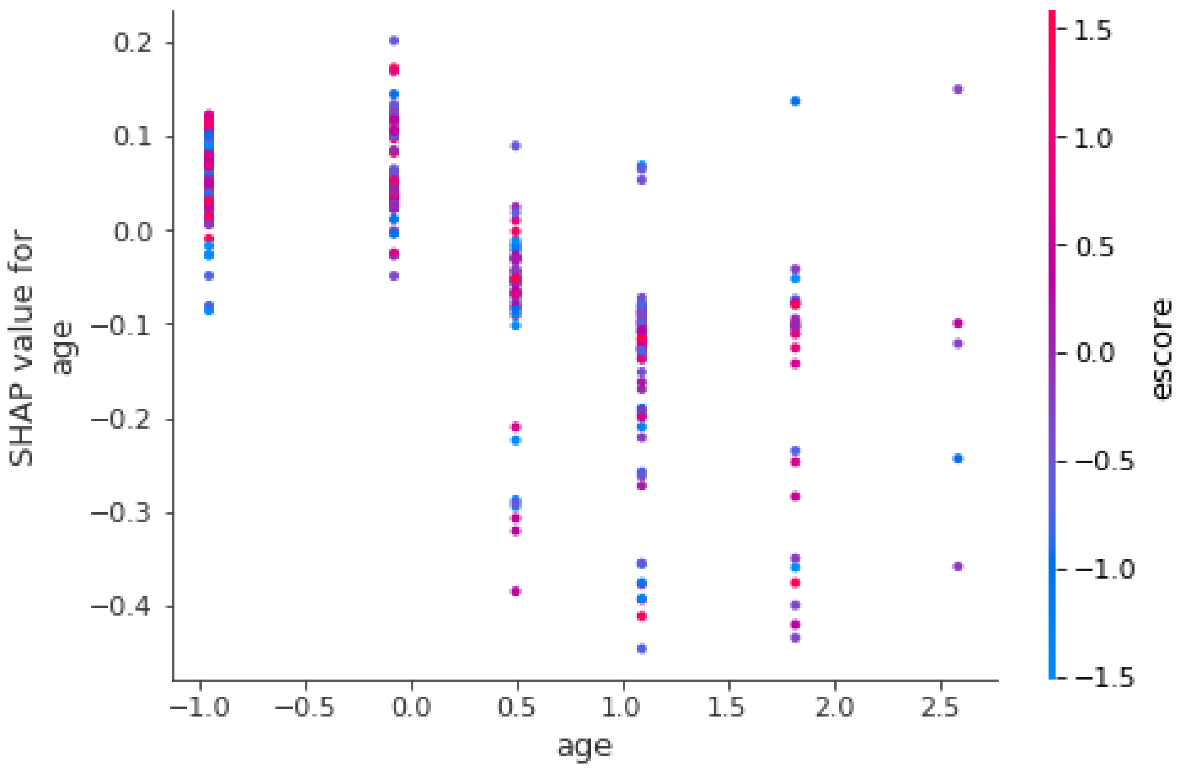



SHAP value for age
age
country
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

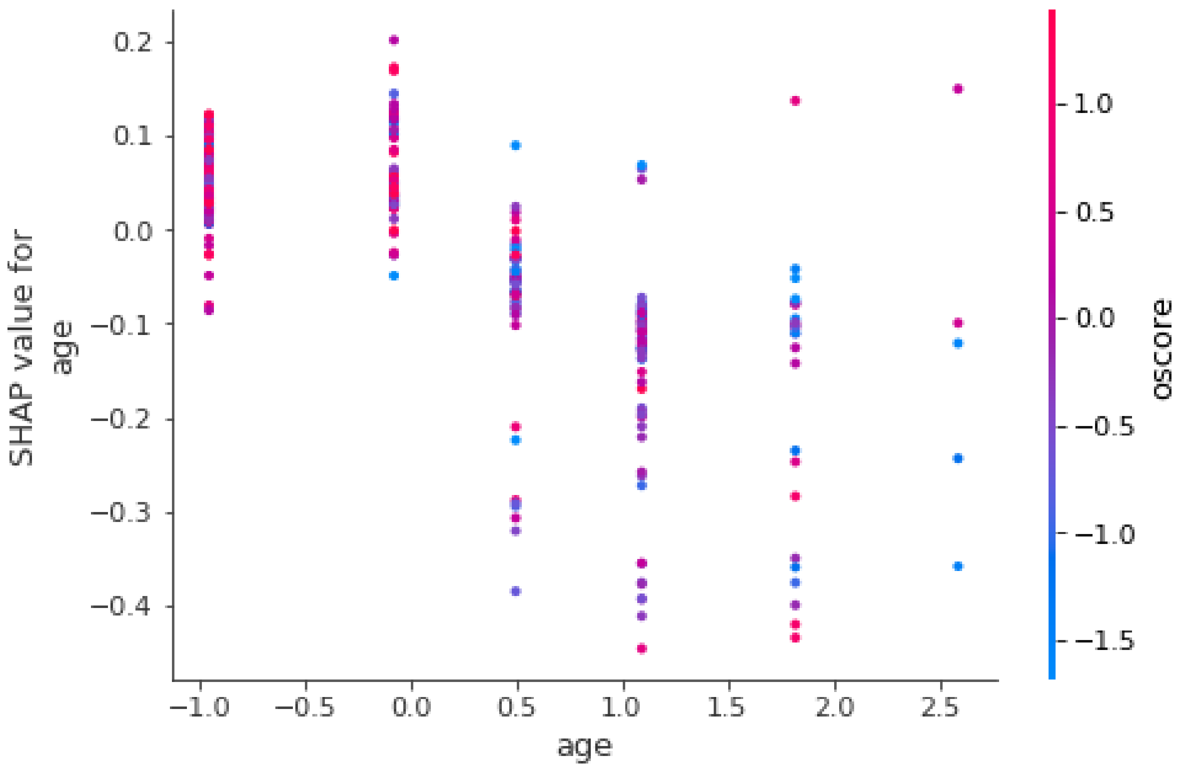

SHAP value for age
age
ethnicity
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
−0.24
−0.26
−0.28
−0.30



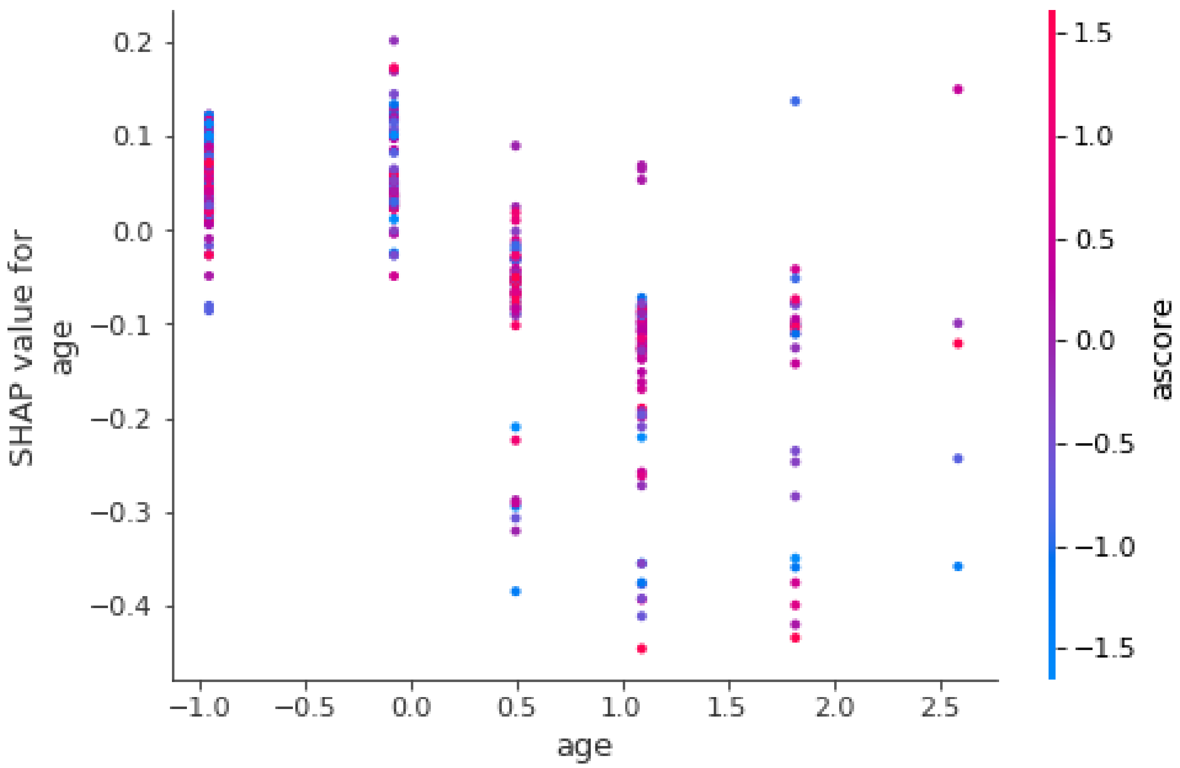
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
nscore



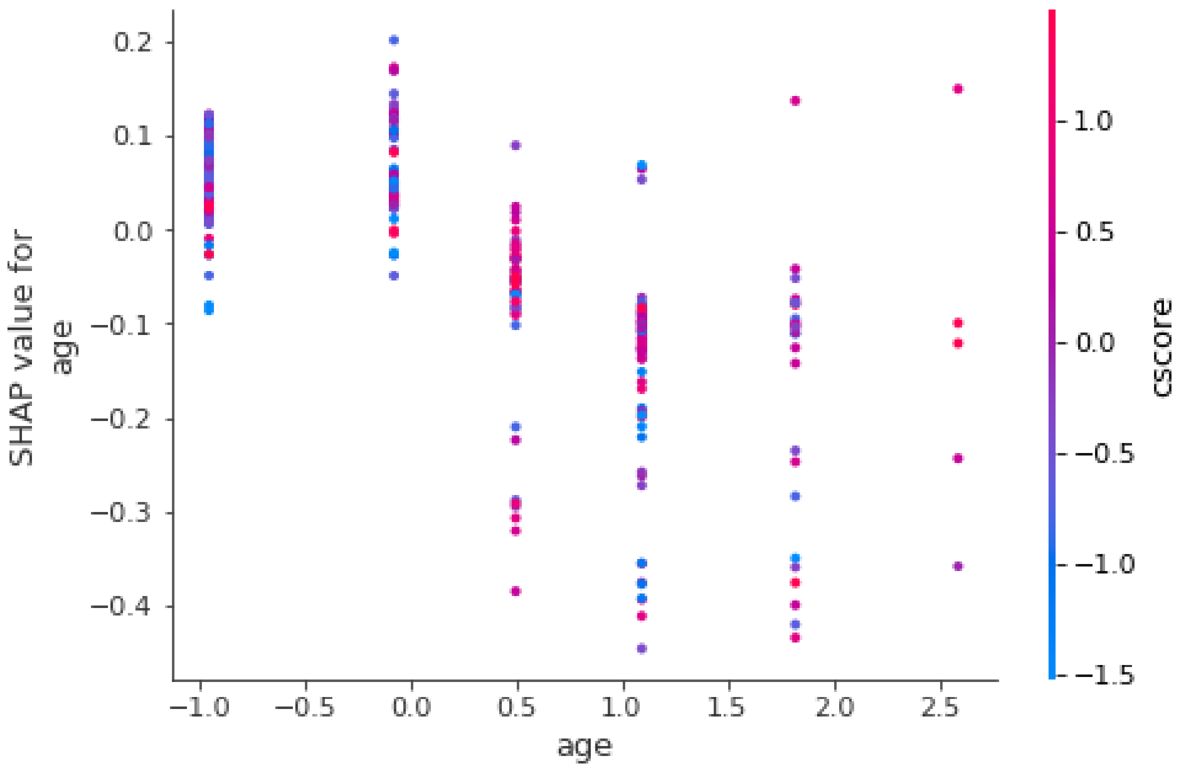
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
SHAP value for age
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
age
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
escore

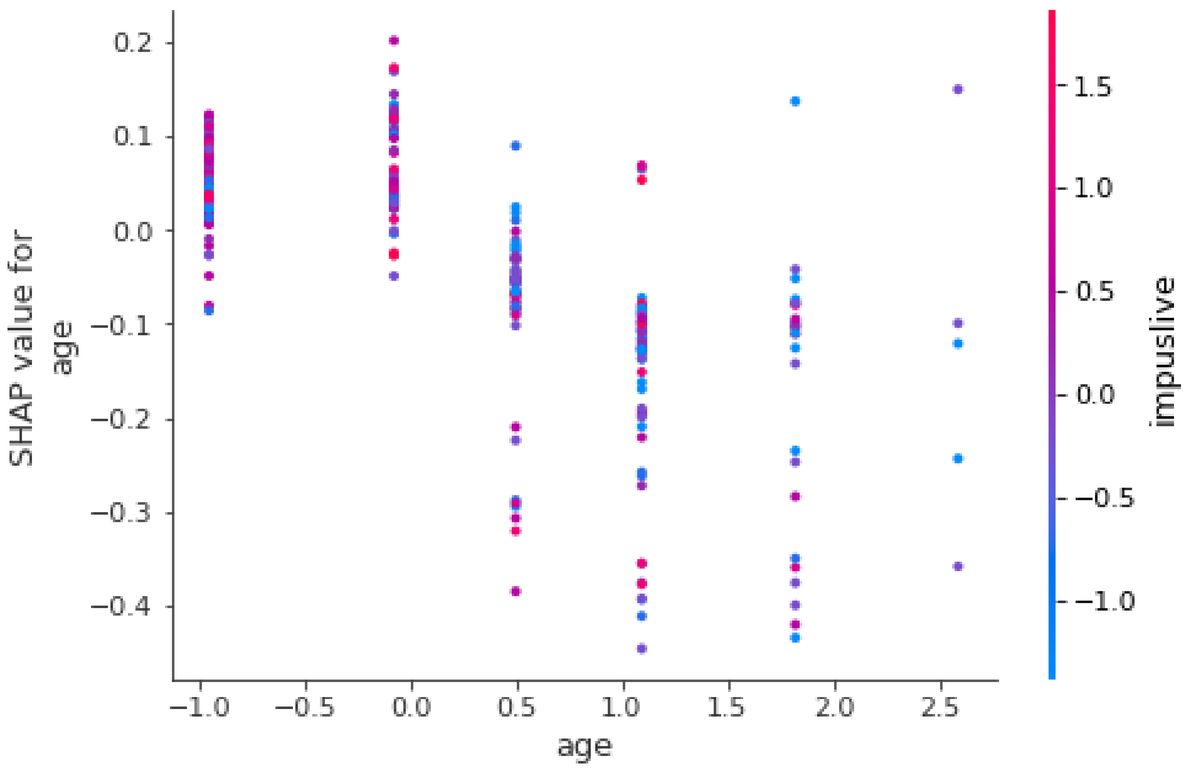


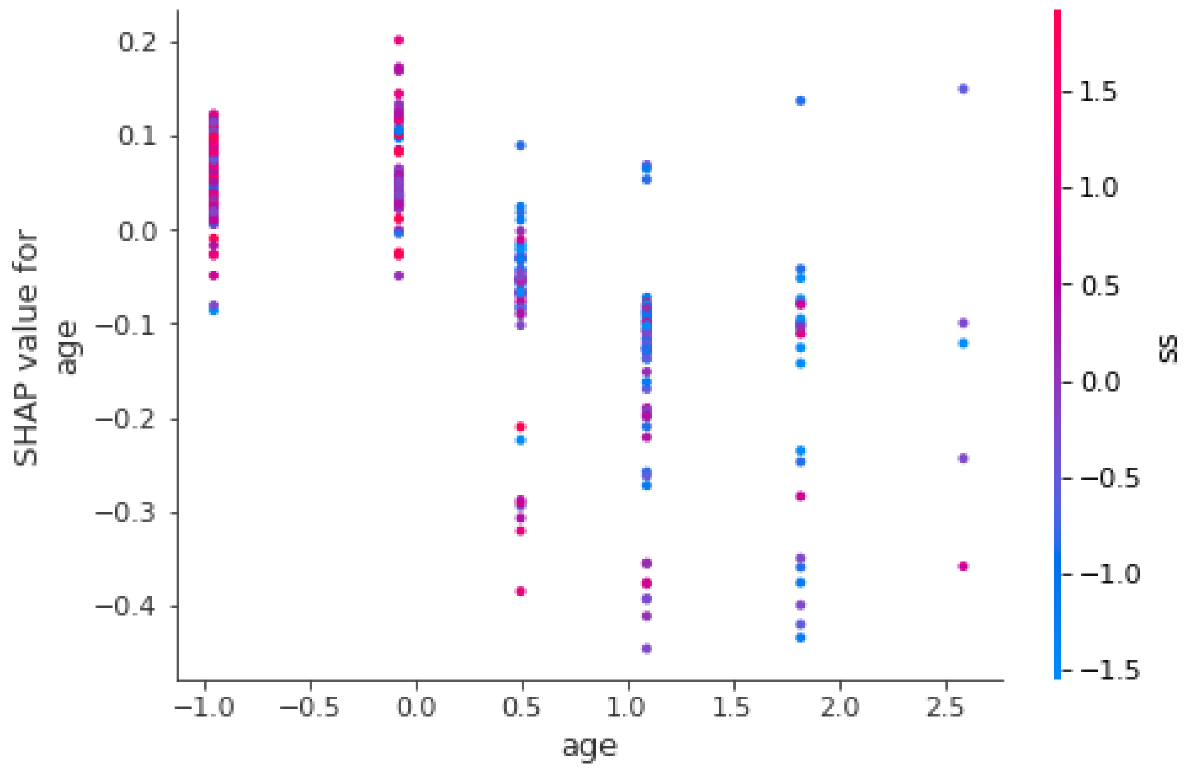
SHAP value for age
age
oscore



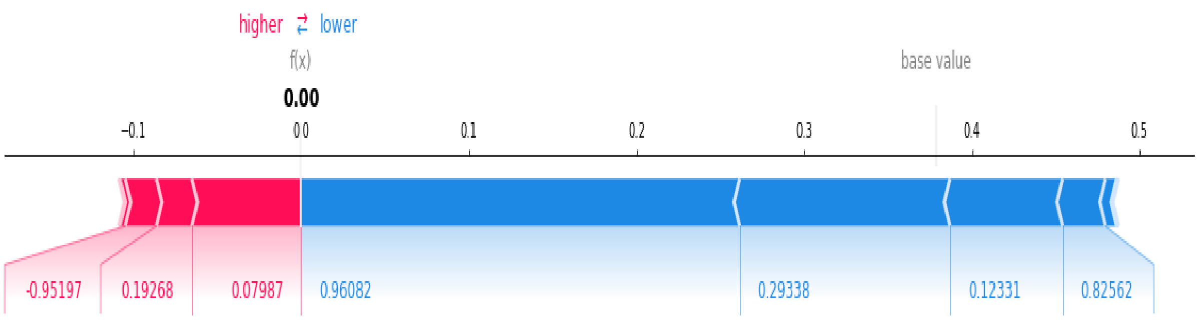
SHAP value for age
age
ascore


SHAP value for age
age
cscore
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

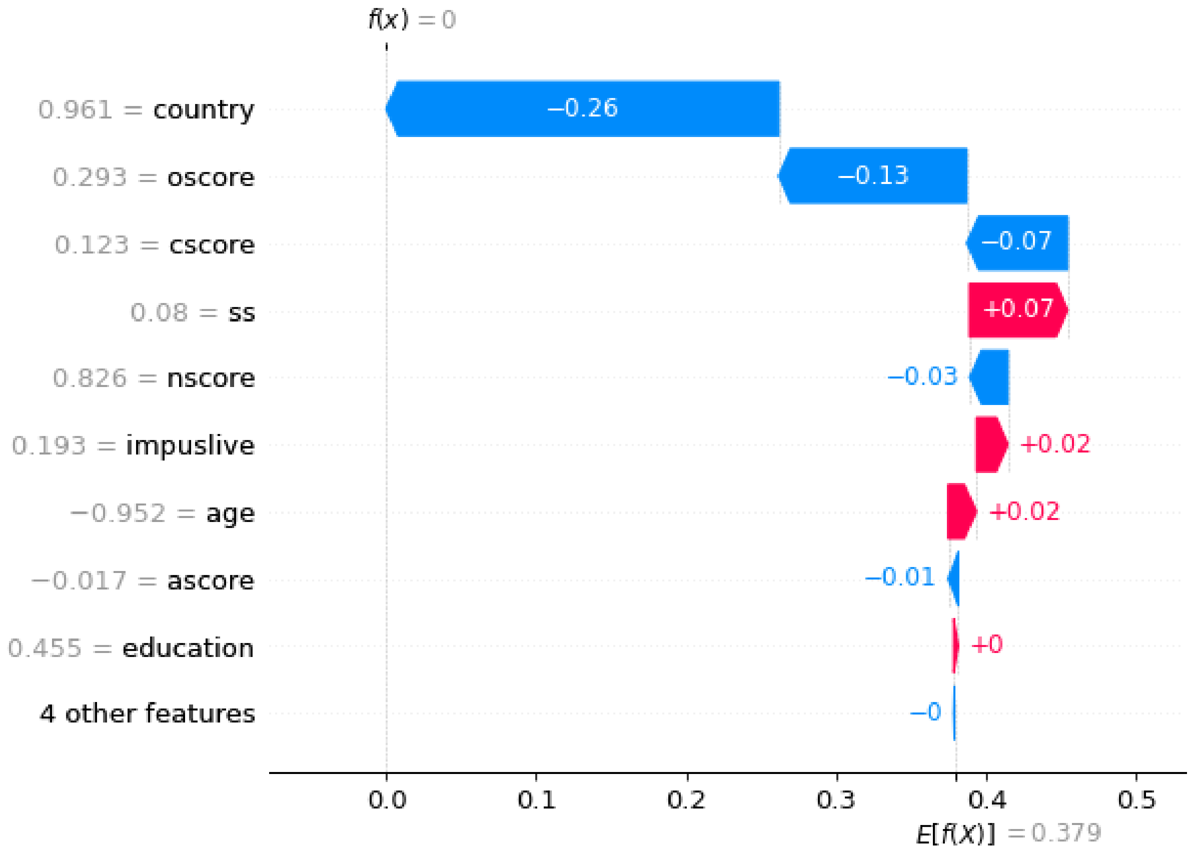

SHAP value for age
age
impuslive
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
1.5
1.0
0.5
0.0
−0.5
−1.0



SHAP Force Plot for a single prediction in Model Decision Tree



SHAP Waterfall Plot for a single prediction in Model Decision Tree



SHAP Decision Plot for Model Decision Tree on Dataset Mushroom Test Set