Fair ML - Homework #2

click here for the CODE FILE

1. Information on features on dataset

	feature	min	max	data_type
0	Age	17	90	float32
1	Workclass	0	8	int8
2	Education-Num	1	16	float32
3	Marital Status	0	6	int8
4	Occupation	0	14	int8
5	Relationship	0	5	int64
6	Race	0	4	int8
7	Sex	0	1	int8
8	Capital Gain	0	99999	float32
9	Capital Loss	0	4356	float32
10	Hours per week	1	99	float32
11	Country	0	41	int8

Description of the features:

Feature	Type	Description
Age	Continuous (float)	Age of the individual in years.
Workclass	Categorical	Type of employment (e.g., Private, Self-emp, Government).
Education-Num	Continuous (float)	Numeric encoding of the education level.
Marital Status	Categorical	Current marital status (e.g., Married, Never-married, Divorced).
Occupation	Categorical	Type of occupation (e.g., Exec-managerial, Sales, Tech-support).
Relationship	Categorical	Relationship role in the household (e.g., Husband, Not-in-family).
Race	Categorical	Ethnicity of the individual.

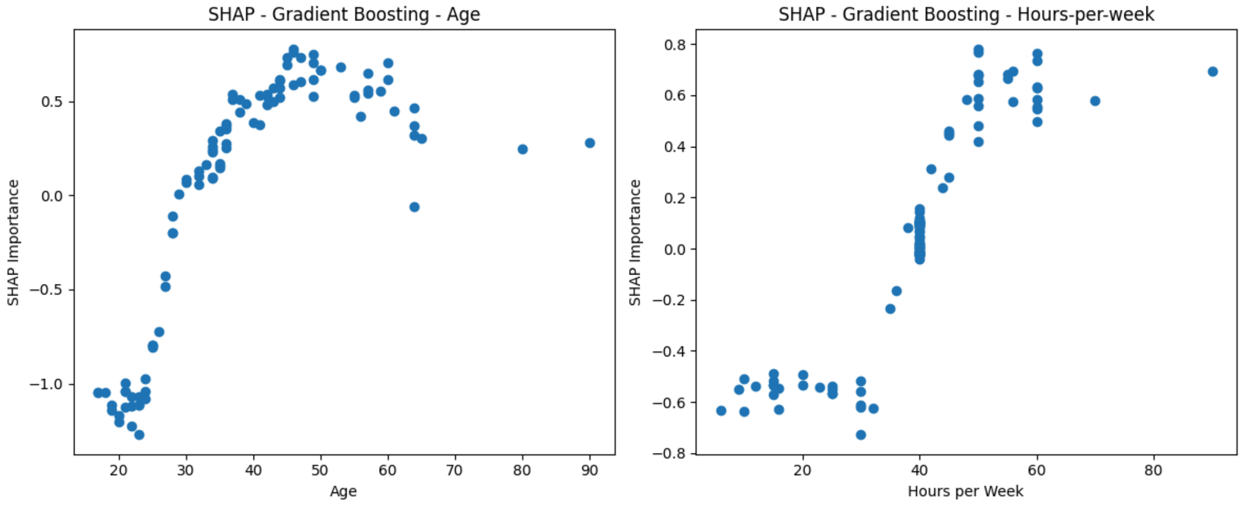
Sex	Categorical	Gender of the individual.
Capital Gain	Continuous (float)	Amount of capital gain received.
Capital Loss	Continuous (float)	Amount of capital loss incurred.
Hours per week	Continuous (float)	Average number of hours worked per week.
Country	Categorical	Country of origin or residence.

- The target variable is income.
  The prediction task is to determine whether an individual's income is:
  1 → greater than $50K
  0 → less than or equal to $50K

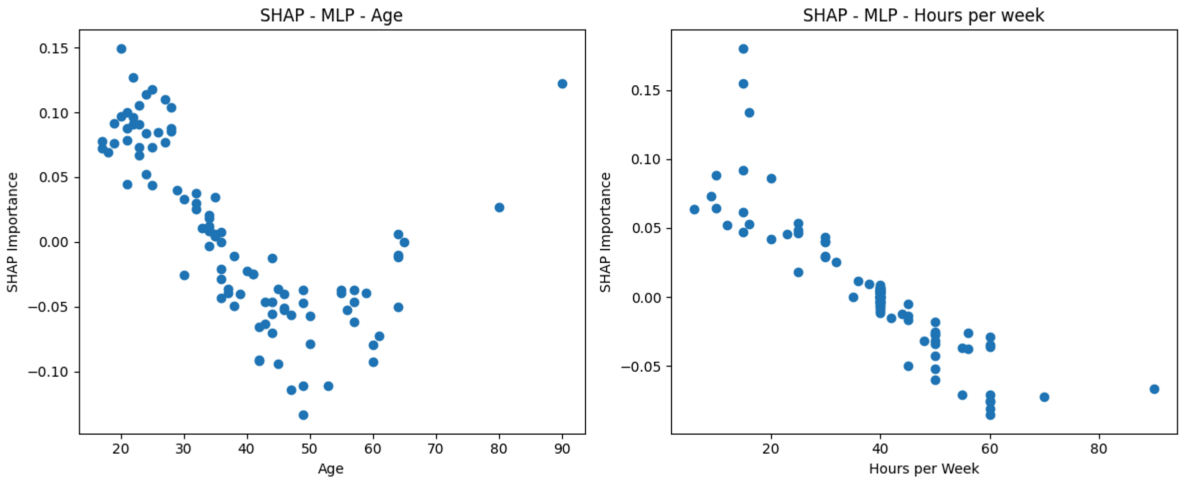
2. Gradient Boosting Train Accuracy: 0.8693565724815725
Gradient Boosting Test Accuracy: 0.8685705512052817

MLP Train Accuracy: 0.8369548525798526
MLP Test Accuracy: 0.8395516658989712

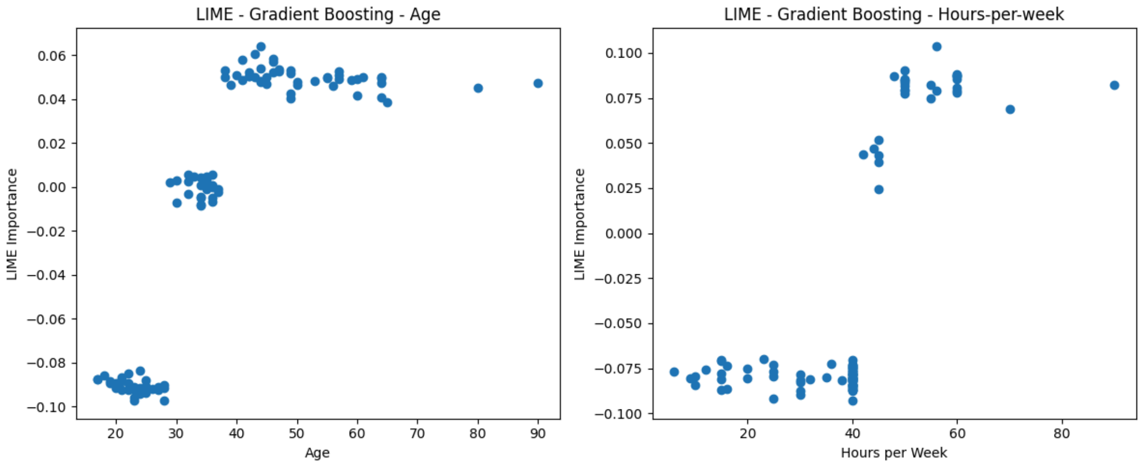
3.



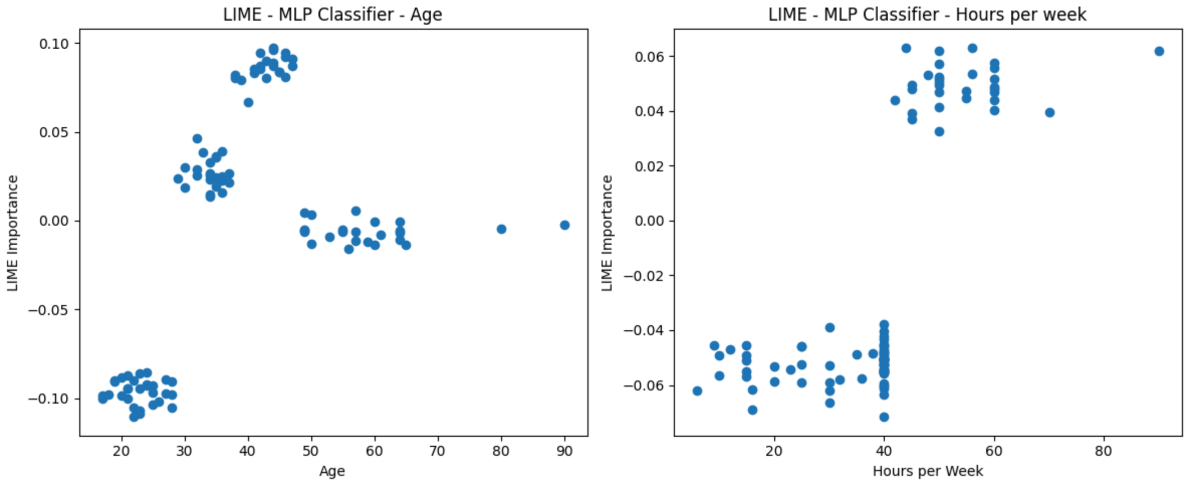
SHAP for Gradient Boosting



SHAP for MLP



LIME for Gradient Boosting



LIME for MLP

- Based on the plots we can see that Age and Hours per week relate to the model predictions using SHAP and LIME.

  SHAP - Gradient Boosting:
    - For Age, feature importance increases till around 50 years of age and then it slightly decreases telling that middle aged individuals are more likely to earn more than 50k.
    - For Hours per week, it shows that importance of more hours worked telling that strong and positive influence on high - income predictions.
- SHAP - MLP Classifier:
    - For Age, there is a negative correlation showing younger individuals have higher SHAP importance and older individuals contribute less positive.
    - For Hours per week, gives decreasing trend in importance which is little inverse to GB possibly due to different internal feature interactions in neural network.
- LIME - Gradient Boosting:
    - For Age, younger and older groups show different levels of contribution.
    - For Hours per week, samples above 45 hours per week tend to contribute positively and those below contribute negatively.
- LIME - MLP Classifier:
    - For Age, it shows that similar behaviour with some groups having strong positive or negative influence.
    - For Hours per week, more hours give positive importance and lower hours show negative.

- If we look at how these features relate to model predictions then,
    - For Age,
        - In GB model, SHAP values shows that middle age people contribute more positive toward predicting >50k, while younger and older individual contributes less.
        - In MLP model, age shows negative trend while the older individuals tend to decrease the model confidence in predicting the high income.
    - For Hours per week,
        - For both SHAP and LIME, individual working with more hours per week are generally associated to positive contributions predicting income>50k
        - Working more than 45-50 hours a week consistently increases prediction confidence in both models while low hours contribute negatively.

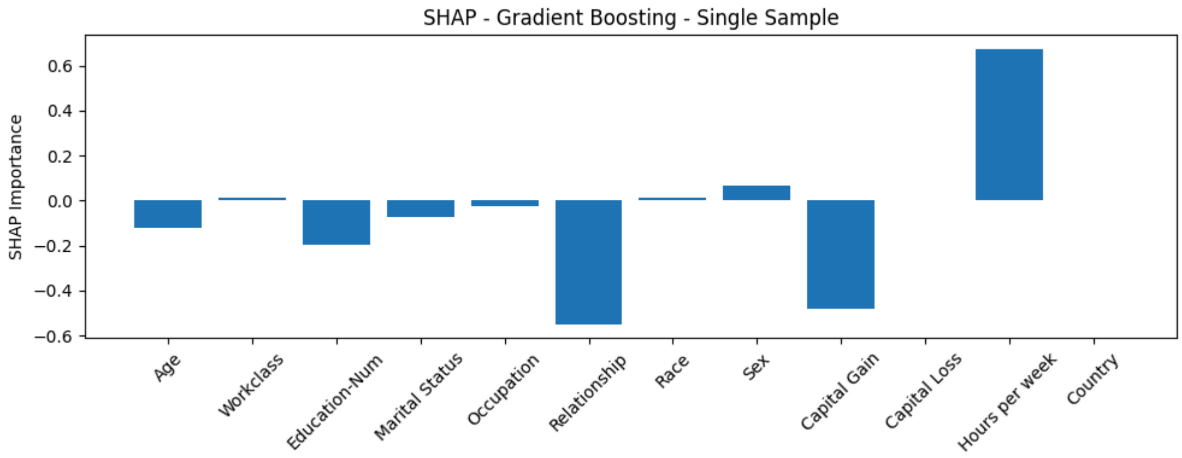
(a.) Yes, the features importances vary,

- In Gradient Boosting, Hours per week consistently shows strong positive importance and Age has peak influence around mid age.
- In MLP, patterns are more distributed. SHAP importance for Age decreases with age, and Hours per week appears less influential overall, showing model has learned different internal representations.

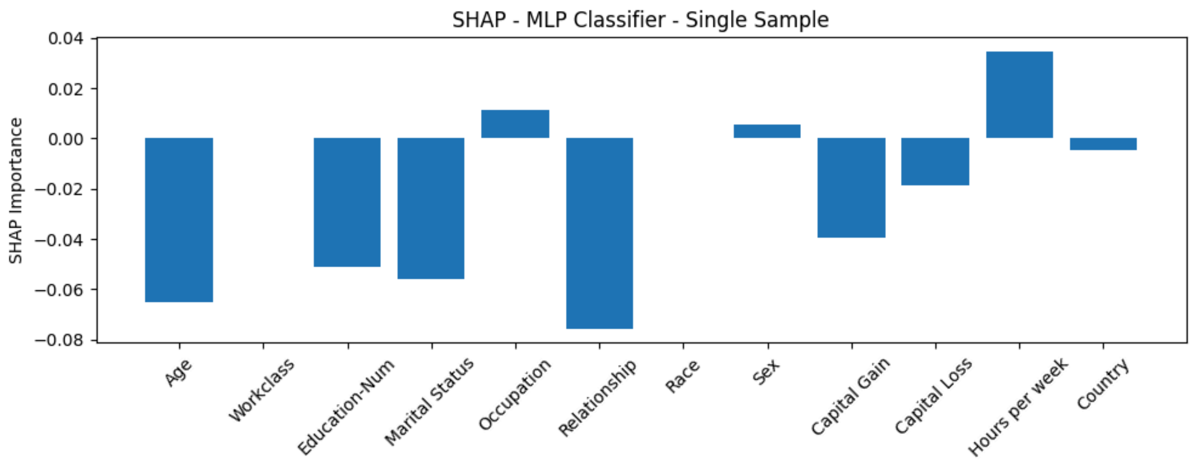
(b.) Yes, feature importances vary across interpretability methods,

- SHAP gives smooth, continuous importance values and captures non linear trends well.
- LIME explanations are locally linear and often convert continuous features into value ranges or intervals. This means that similar values are grouped together, and assigned the same importance. As a result, feature importance can change suddenly between intervals, leading to more abrupt shifts in interpretation and greater variability across different samples.

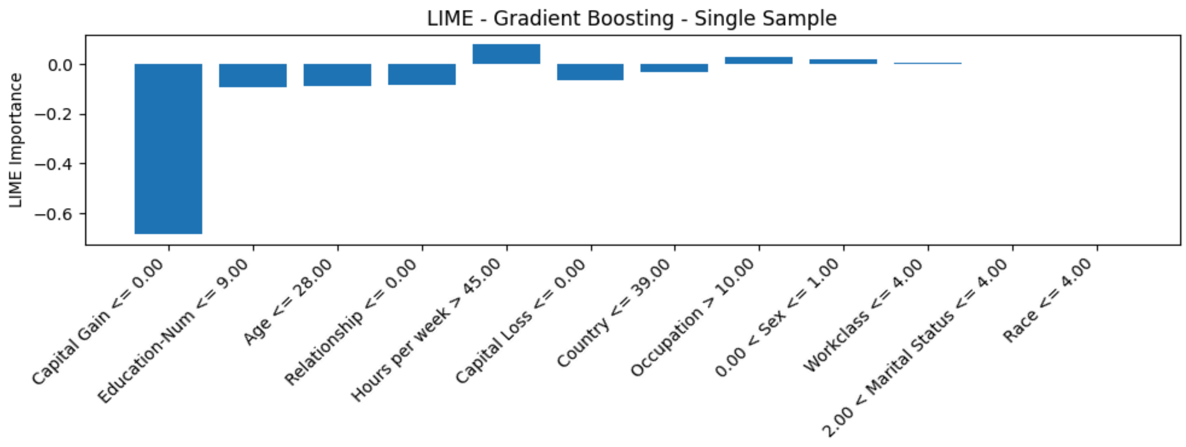
4.



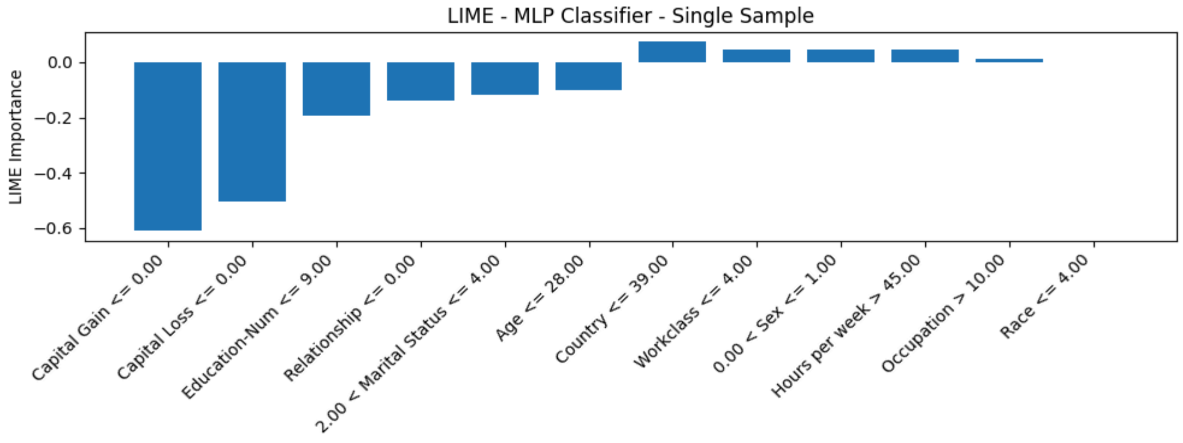
SHAP for GB - single sample



SHAP for MLP - single sample



LIME for GB - single sample



LIME for MLP - single sample

- SHAP – Gradient Boosting
    - Positive Contributor: Hours per week
    - Negative Contributor: Relationship

- SHAP – MLP Classifier
    - Positive Contributor: Hours per week
    - Negative Contributor: Relationship

- LIME – Gradient Boosting
    - Positive Contributor: Hours per week
    - Negative Contributor: Capital Gain

- LIME – MLP Classifier
    - Positive Contributor: Country
    - Negative Contributor: Capital Gain

For the selected sample, we analyzed SHAP and LIME explanations for both Gradient Boosting and MLP classifiers. The most consistent positive contributor across all methods was hours per week, reflecting the model's reliance on working hours as a signal for income. In contrast, relationship and capital gain repeatedly appeared as negative contributors, especially in SHAP and LIME for Gradient Boosting.

Interestingly, in LIME for MLP, country, emerged as the top positive contributor, revealing that local interpretability can uncover different influencing features compared to global trends.