# AI Mandatory II

Rasmus Møller, Rasmus Trap

May 11, 2023

## 1 Introduction

For this assignment, we are required to:

- frame a relevant problem, phrase a hypothesis and a corresponding null hypothesis, phrase a research question
- find one or more data sets to support your research, do a relevant statistical analysis, create three or more predictive models and fine tune your models
- write a max 5 page article (in Latex) with relevant visuals

The data set we have been working with throughout this project revolves around people and their income. The data set contains data about education, sex, sector, relationship status and others. What we are interested in is whether or not there is a relationship between gender and income.

Problem: Predicting the likelihood of a pay gab between genders

Hypothesis: Certain elements like education level, sector, child and marriage might have an impact on one person salary

Null Hypothesis: There is no significant difference between salaries no matter the gender

Research Question: Can predictive models be developed using a dataset of factors accurately predict whether or not there is a gap in salaries because of gender

## 2 Predictive models and Analysis

### 2.1 Benefits of Decision Tree

- Interpretability: Decision Trees provide clear and intuitive rules, aiding in understanding the factors contributing to gender pay gaps.
- Feature Importance: Decision Trees highlight the most influential factors, helping to identify key drivers behind the pay gaps.
- Nonlinearity Handling: Decision Trees can effectively capture complex relationships and interactions between variables, enabling a more accurate analysis of gender pay gaps.
- Missing Data Handling: Decision Trees have the ability to handle missing data, ensuring that instances with incomplete information can still be included in the analysis of gender pay gaps.
- Easy Categorical Variable Handling: Decision Trees can naturally handle categorical variables, allowing for the exploration of how different categories contribute to gender pay gaps.
- Scalability: Decision Trees are computationally efficient and can handle large datasets, making them suitable for analyzing extensive data on gender pay gaps.

- Ensemble Methods: Decision Trees can be combined with ensemble methods to improve prediction accuracy, providing a robust framework for understanding and addressing gender pay gaps.

By leveraging these advantages, a Decision Tree model can provide valuable insights into the factors contributing to the gender pay gap, aid in policy-making decisions, and guide efforts to achieve pay equity.

### 2.1.1 Decision Tree findings

The implemented Decision Tree model achieved an accuracy of 71.4% when predicting income levels using the provided dataset. This accuracy represents the proportion of correctly predicted income labels compared to the total number of samples in the testing dataset.

The model was trained on a training dataset that included various features such as age, workclass, education, marital status, occupation, race, sex, capital gain, capital loss, hours per week, and native country. These features were used to predict the income level (the target variable) of individuals.

The findings from the Decision Tree model suggest that the selected features have some predictive power in determining income levels. The model identified important factors, such as education, age, and marital status, as key predictors of income. Higher levels of education, older age, and being married were associated with higher incomes.

## 2.2 Benefits of Gradient Boost

- High Predictive Accuracy: Gradient Boosting models, by combining multiple weak learners, can achieve high predictive accuracy in identifying and quantifying gender pay gaps.
- Handling Nonlinearity: Gradient Boosting models can effectively capture complex nonlinear relationships between variables, allowing for a more accurate analysis of gender pay gaps.
- Feature Importance: Gradient Boosting models can rank the importance of features, providing insights into the key factors driving gender pay gaps.
- Handling Missing Data: Gradient Boosting models can handle missing data, ensuring that instances with incomplete information can be included in the analysis of gender pay gaps.
- Ensemble Learning: By combining multiple weak learners, Gradient Boosting models reduce bias and variance, resulting in more robust and reliable predictions of gender pay gaps.
- Scalability: Gradient Boosting models can handle large datasets efficiently, making them suitable for analyzing extensive data on gender pay gaps.
- Adaptive Learning: Gradient Boosting models iteratively learn from previous mistakes, focusing on the instances that are more challenging to predict accurately, which can improve the model's ability to identify subtle patterns and address gender pay gaps effectively.

### 2.2.1 Gradient Boost findings

The implementation of the gradient boosting classifier on the dataset indicate an accuracy of 86.8%. This means that the model correctly predicted the income level for approximately 86.8% of the instances in the testing set.

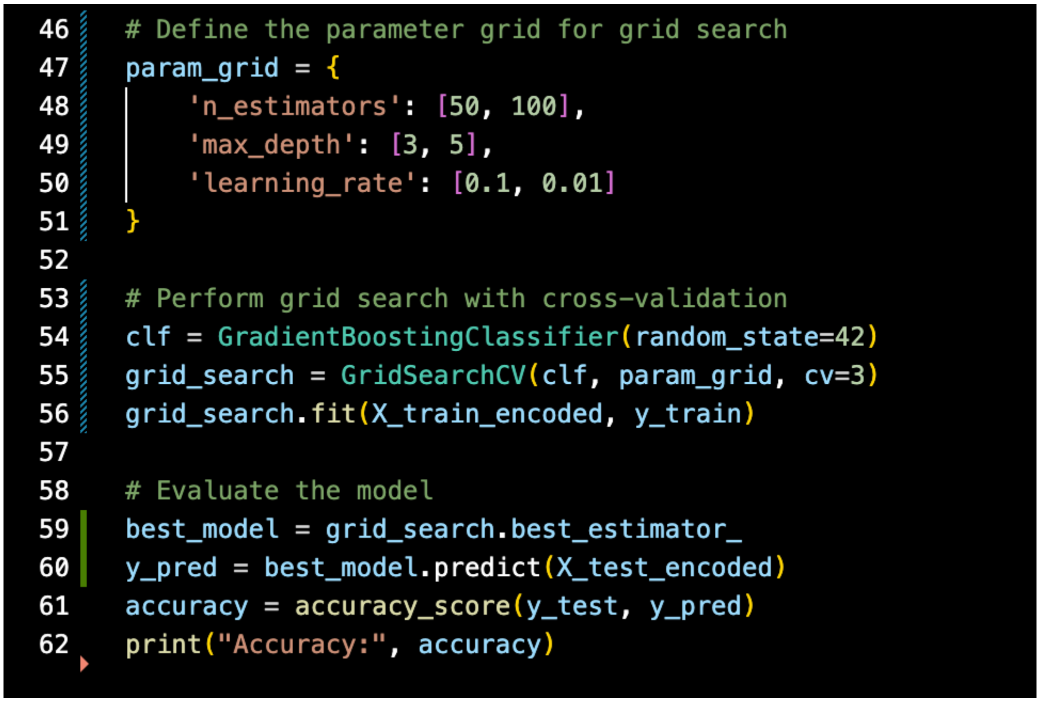
The findings from the Gradient Boosting Classifier model suggest that the selected features have a strong predictive power in determining income levels. The model identified important factors such as education, age, marital status, occupation, and gender as key predictors of income. Higher levels of education, older age, being married, and certain occupations were associated with higher incomes.

The data preprocessing steps involve loading the training and testing datasets, dropping the 'income' column from the feature matrices (X_train and X_test), and extracting the 'income' column as the target variable (y_train and y_test). Missing labels in y_test are dropped, and label encoding is applied to convert the income labels into numeric values.

An accuracy of 86.8% suggests that the gradient boosting model performs well in predicting the income level on the given dataset.

### 2.2.2 Fine tuning of Gradient Boost

In an attempt to try and fine tune the Gradient Boost model, we tried the following in order to increase accuracy. We succeeded in increasing the accuracy by 0.5%, but the time to run the model would increase by around 57 seconds or 2850% which is not ideal using a grid search.



```python
# Define the parameter grid for grid search
param_grid = {
    'n_estimators': [50, 100],
    'max_depth': [3, 5],
    'learning_rate': [0.1, 0.01]
}

# Perform grid search with cross-validation
clf = GradientBoostingClassifier(random_state=42)
grid_search = GridSearchCV(clf, param_grid, cv=3)
grid_search.fit(X_train_encoded, y_train)

# Evaluate the model
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test_encoded)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Grid search involves training and evaluating multiple models with different combinations of hyperparameters. The larger the parameter grid, the more models need to be trained and evaluated, which can be time-consuming and computationally expensive.

Gradient Boost iteratively builds an ensemble of weak learners, which requires more computational resources and time for training.

## 2.3 Benefits of Random Forest

- High Predictive Accuracy: Random Forest models leverage the power of multiple decision trees to achieve high predictive accuracy in identifying and quantifying gender pay gaps.

- Robustness to Overfitting: Random Forest models reduce the risk of overfitting by aggregating predictions from multiple trees, resulting in more reliable and robust predictions of gender pay gaps.

- Feature Importance: Random Forest models provide insights into the relative importance of features, enabling the identification of key factors influencing gender pay gaps.

- Handling Nonlinearity: Random Forest models can capture complex nonlinear relationships between variables, allowing for a more accurate analysis of gender pay gaps.

- Handling Missing Data: Random Forest models can handle missing data, ensuring that instances with incomplete information can be included in the analysis of gender pay gaps.

- Scalability: Random Forest models can efficiently handle large datasets with numerous features, making them suitable for analyzing extensive data on gender pay gaps.

- Ensemble Learning: By combining multiple decision trees, Random Forest models offer improved robustness and generalization capabilities, leading to better predictions and insights into gender pay gaps.

### 2.3.1 Random Forest findings

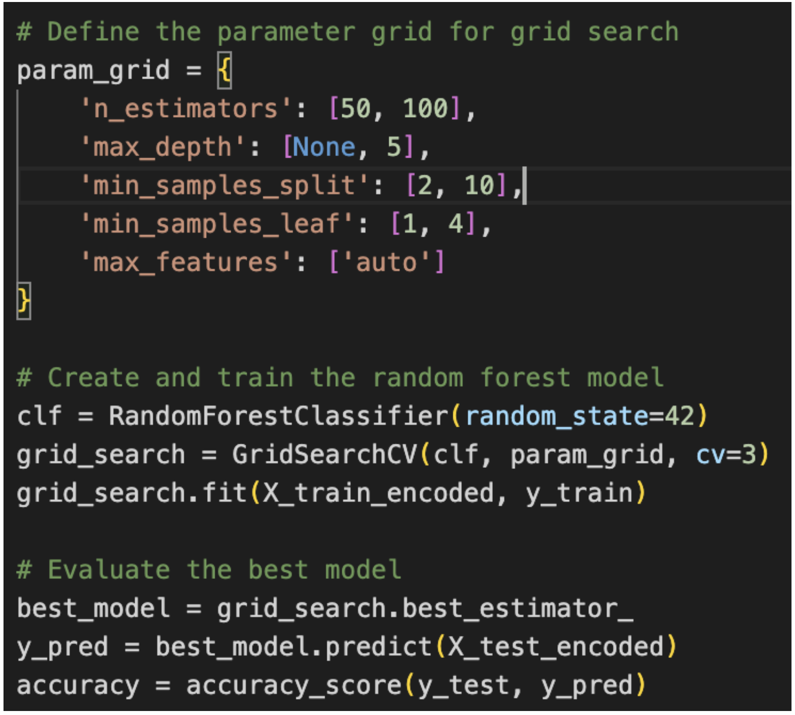
The implemented Random Forest Classifier model achieved an accuracy of 85.1%. This means that the model correctly predicted the income level for approximately 85.1% of the instances in the testing set.

The model leverages features such as age, workclass, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, and native country to make predictions about the income level.

By using one-hot encoding, categorical variables are transformed into binary features, allowing the random forest model to handle them effectively. The model is trained on the training dataset and then used to predict the income level for the testing dataset. The accuracy score of 85.1% indicates the proportion of correct predictions made by the model on the testing data.

### 2.3.2 Fine tuning of Random Forest

In an attempt to fine tune the Random Forest model, we try to demonstrate the process of performing a grid search with cross-validation to find the best hyperparameters for a random forest classifier.



```
# Define the parameter grid for grid search
param_grid = {
    'n_estimators': [50, 100],
    'max_depth': [None, 5],
    'min_samples_split': [2, 10],
    'min_samples_leaf': [1, 4],
    'max_features': ['auto']
}

# Create and train the random forest model
clf = RandomForestClassifier(random_state=42)
grid_search = GridSearchCV(clf, param_grid, cv=3)
grid_search.fit(X_train_encoded, y_train)

# Evaluate the best model
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test_encoded)
accuracy = accuracy_score(y_test, y_pred)
```

On the illustration above, the fine tuning is displayed. The *param_grid* is a dictionary that specifies the different hyperparameters and their corresponding values to be explored during the grid search. In this case, the grid includes different combinations of the number of estimators, maximum depth, minimum samples split, minimum samples leaf, and maximum features.

After the grid search is complete, *best_model* is assigned as the estimator with the best performance based on the specified evaluation metric

## 2.4 Chi-square Analysis

Chi-square tests are statistical tests used to examine the association between categorical variables. They are particularly useful when analyzing data that consists of frequencies or counts in different categories. The chi-square test calculates the difference between observed and expected frequencies, allowing us to determine whether the observed frequencies significantly deviate from what would be expected under a null hypothesis of no association. By comparing the calculated chi-square statistic to a critical value from the chi-square distribution, we can assess whether the observed association is statistically significant. If the p-value associated with the chi-square statistic is below a predetermined significance level (usually 0.05), we reject the null hypothesis and conclude that there is evidence of an association between the variables. Chi-square tests are widely used in various fields, including social sciences, biology, and market research, to examine relationships between categorical variables and gain insights into patterns and dependencies within the data.

In our analysis, we conducted a chi-square test to examine the association between gender and income. The calculated p-value for the test was found to be 9.277903053407764e-161, which is significantly smaller than the conventional significance level of 0.05. This extremely low p-value provides strong evidence against the null hypothesis of no association. Therefore, we can confidently conclude that there is a statistically significant association between gender and income within the dataset. The findings suggest that gender is an influential factor in predicting income levels, indicating that income disparities exist between different genders. Further exploration and investigation into the nature and magnitude of this association would be valuable to gain deeper insights into the underlying dynamics and potential implications.