EMBEDDED METHODS

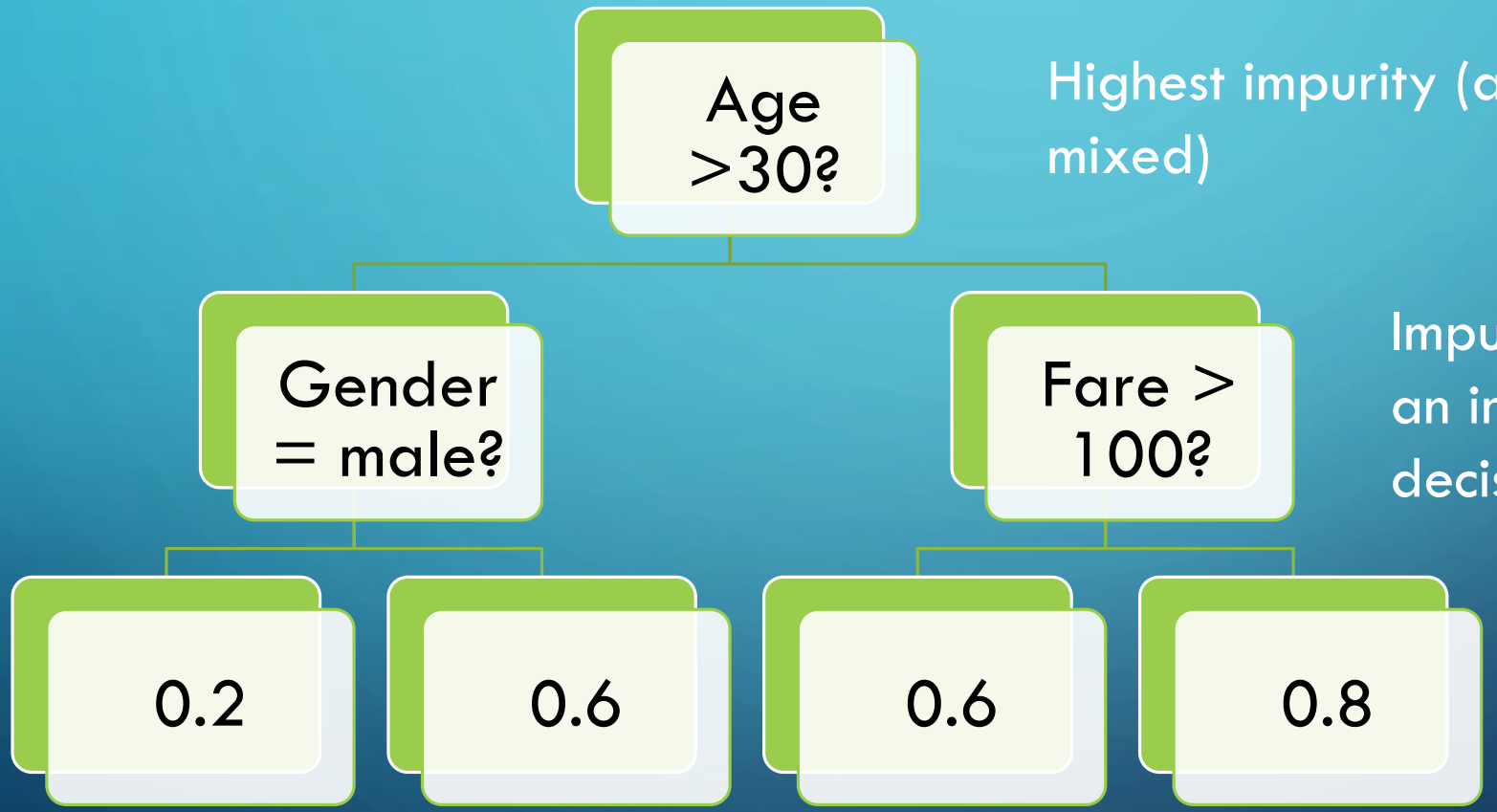
FEATURE SELECTION BY TREE DERIVED VARIABLE IMPORTANCE

DECISION TREES

Decision trees

- Most popular machine learning algorithms
- Highly accurate
- Good generalisation (low overfitting)
- Interpretability

DECISION TREE IMPORTANCE



Highest impurity (all classes are mixed)

Impurity decreases as age is an important factor in the decision

Impurity decreases further

RANDOM FOREST IMPORTANCE

- Random Forests consist of several hundreds of individual decision trees
- The impurity decrease for each feature is averaged across trees

Limitations

- Correlated features show equal or similar importance
- Correlated features importance is lower than the real importance, determined when tree is built in absence of correlated counterparts
- Highly cardinal variables show greater importance (trees are biased to this type of variables)

RANDOM FOREST IMPORTANCE

- Build a random forest
- Determine feature importance
- Select the features with highest importance
- There is a scikit-learn implementation for this

RANDOM FOREST IMPORTANCE

Recursive feature elimination

- Build random forests
- Calculate feature importance
- Remove least important feature
- Repeat till a condition is met

- If the feature removed is correlated to another feature in the dataset, by removing the correlated feature, the true importance of the other feature will be revealed → its importance will increase

GRADIENT BOOSTED TREES FEATURE IMPORTANCE

- Feature importance calculated in the same way
- Biased to highly cardinal features
- Importance is susceptible to correlated features
- Interpretability of feature importance is not so straightforward:
  - Later trees fit to the errors of the first trees, therefore feature importance is not necessarily proportional on the influence of the feature on the outcome, rather on the mistakes of the previous trees.
  - Averaging across trees may not add much information on true relation between feature and target