Norwegian University of Science and Technology
Department for Computer Science

Methods in Artificial Intelligence
TDT4171
Spring 2021

Jørgen Rosager
Exercise 1

$\boxed{1}$ **a)** The continuous variables are Age, $Name$, $Ticket$, $Cabin$, $Fare$ as these are pretty much unique for almost all (except for $Cabin$ where more people can live in same cabin, but this does not warrant enough discrete cases to make it discrete). Both $SibSp$ and $Parch$ can be treated as both as for most persons this number should be somewhere between 0-10. However, there are edge cases and this is in reality a continuous variable. This means that when we train it with these as discrete attributes there is a chance that the DT meets a unhandled case (for example $SibSp = 11$) when used on testing data. In my DT model I have handled this by making an educating guess if $Parch$ or $SibSp$ are chosen as discrete attributes and there is no known case of this from the training data. This proved to be smart as the highest accuracy was gotten when using $SibSp$ as a discrete attribute (see results below).

When trained on all discrete attributes these were the results:

This model is extremely complicated and got an accuracy of 86.8%.

The attribute *Embarked* shouldn't really affect the survival of the passenger so I decided to drop this. The results:

The graph got a bit simpler and the accuracy increased to 88.0%, this makes sense because the attribute *Embarked* was overfitting the problem.

In fact the best results are gotten by also removing the *Parch* attribute:

The graph is now easily readable the accuracy improved to 88.5%.

b) Support for continuous variables were added by changing the *gain* function to loop over the unique values of the continuous attribute and then find the best split point. The results by using the attributes Age, $Parch$ and $SibSp$ as continuous attributes:

The accuracy was 86.6% which is slighty lower than the results gotten by using discrete values on *Parch* and *SibSp* and without *Age*.

c) As said the DT from **a)** performed better than the DT from **b)**. Personally I believed that using the *Age* attribute would be relevant, especially since they followed the Women and children first code of conduct when saving passengers on the Titanic.

1. Bootstrapping: https://towardsdatascience.com/boosting-the-accuracy-of-your-machine-learning-models-f878d6a2d185 2. Chi squared pruning (feature selection/feature importance)

2 https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4 Column: Cabin missing data...

KNN: Weakness: Need some sort of distance measurement Time expensive Advantages: Simple to understand and implement (still higher than mean/median) "One of the obvious drawbacks of the KNN algorithm is that it becomes time-consuming when analyzing large datasets because it searches for similar instances through the entire dataset. Furthermore, the accuracy of KNN can be severely degraded with high-dimensional data because there is little difference between the nearest and farthest neighbor."

Prediction model: Use regression: Weaknesses: Can be very expensive operation, Requires a lot of domain knowledge to say what should influence the *Cabin* attribute. Advantages: Can be very accurate

Mean, median imputation: Weaknesses: Takes no advantage of relationship between variables Advantages: