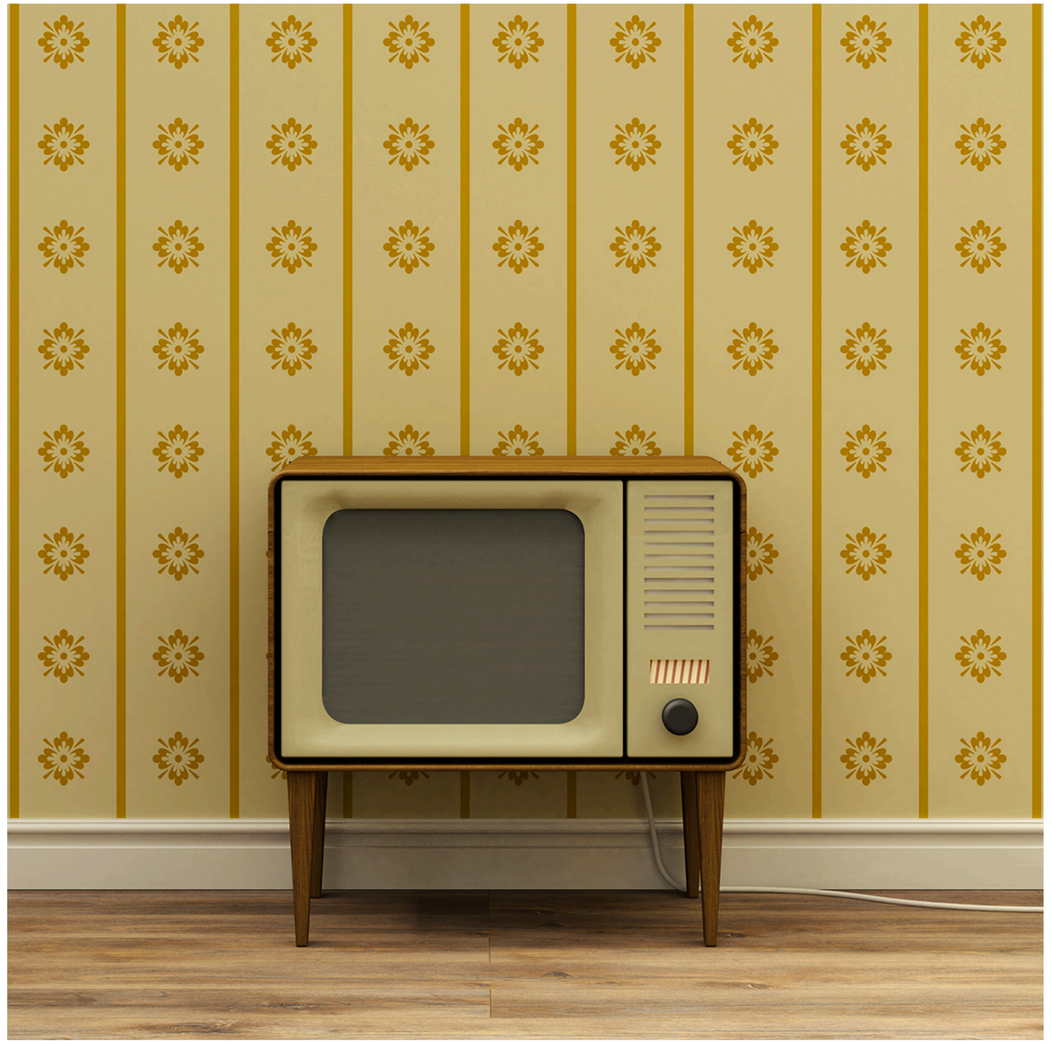
MACHINE LEARNING SESSION 3

Vaibhav Verdhan

Jan 22, 2021

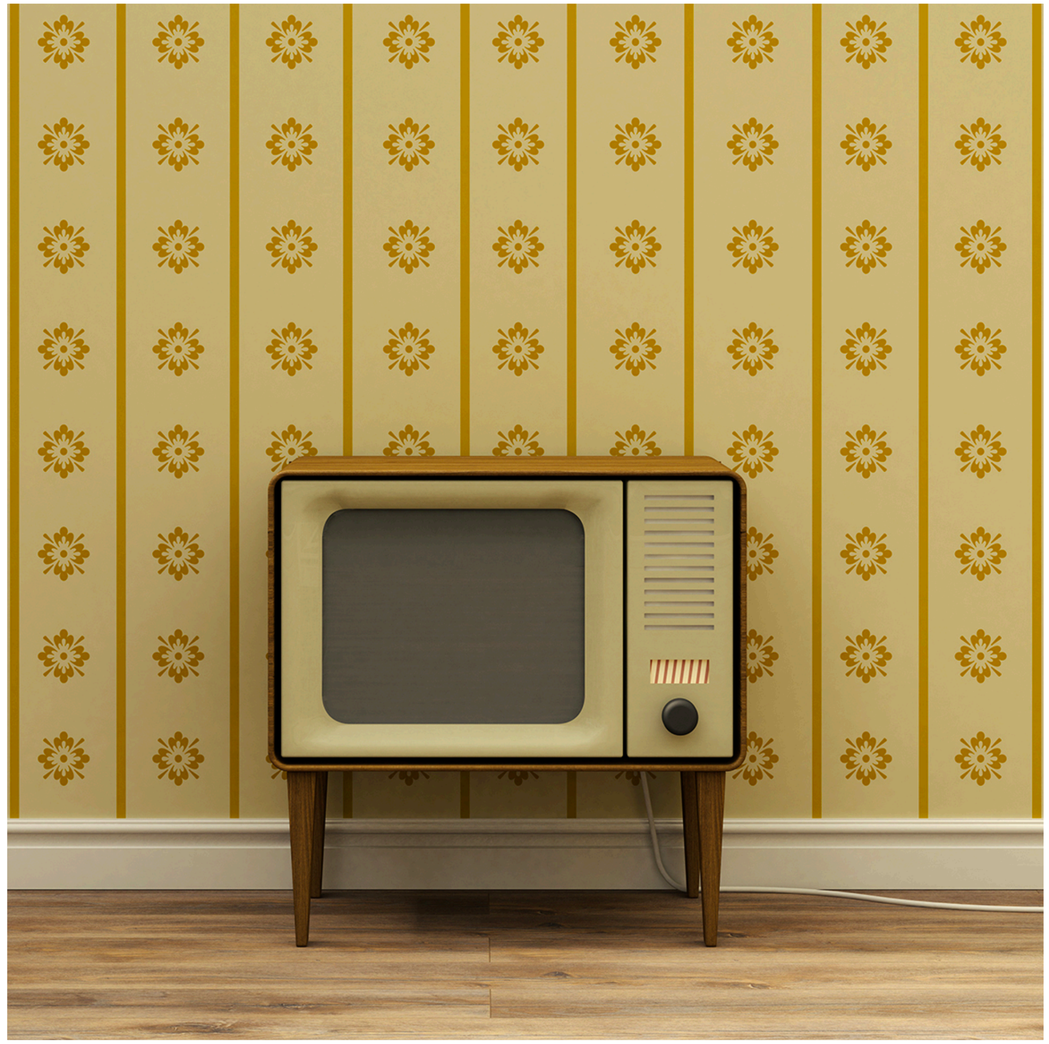
AGENDA FOR THE SESSIONS

Session	Date	Agenda
1	Jan 8, 2021	Introduction to ML
2	Jan 15, 2021	Supervised Learning - 1
3	Jan 22, 2021	Supervised Learning - 2
4	Jan 29, 2021	Unsupervised Learning - 1
5	Feb 05, 201	Unsupervised Learning - 2



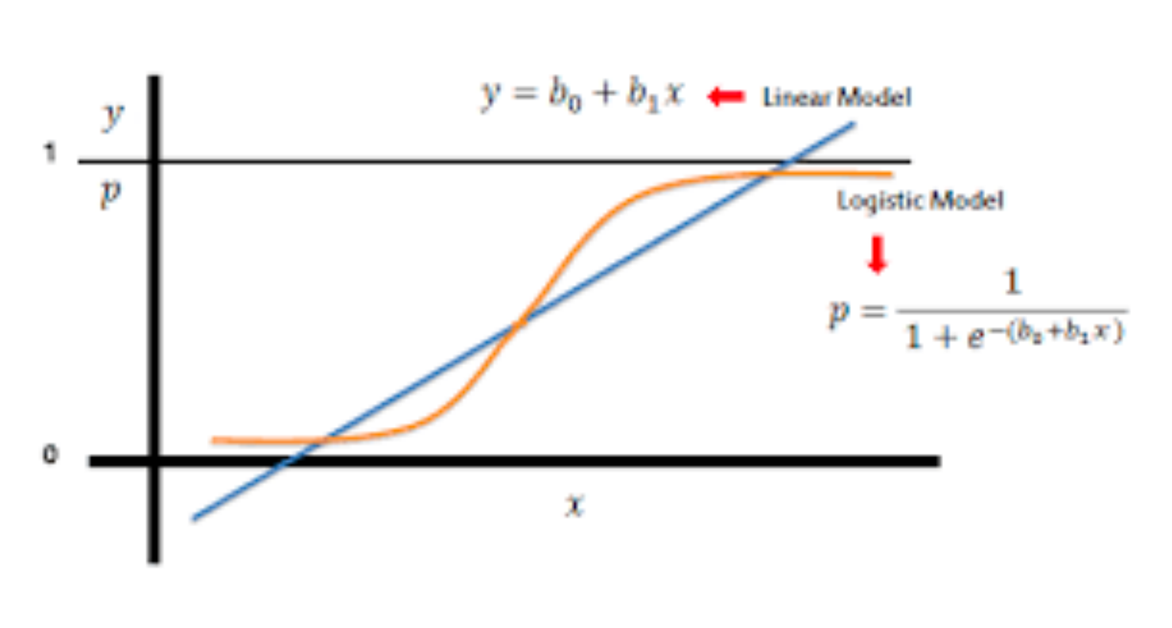
QUESTIONS WE WILL DISCUSS TODAY

- Supervised Learning
- Classification problems
- Use Cases of classification problems
- Logistic Regression
- Decision Tree
- Random Forest
- Confusion Matrix, ROC/AUC
- Python Implementations



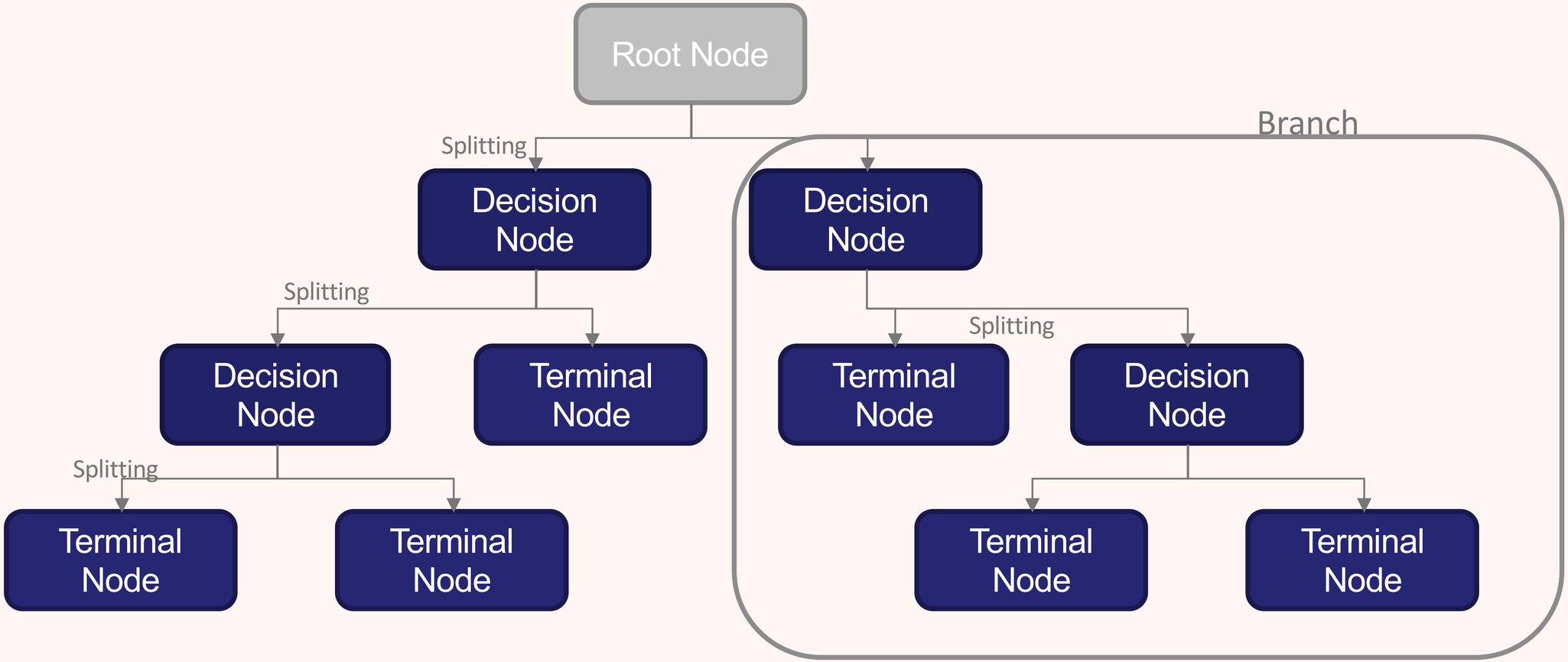
Classification problems are used for a categorical variable

Logistic regression is generally the first algo used



$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

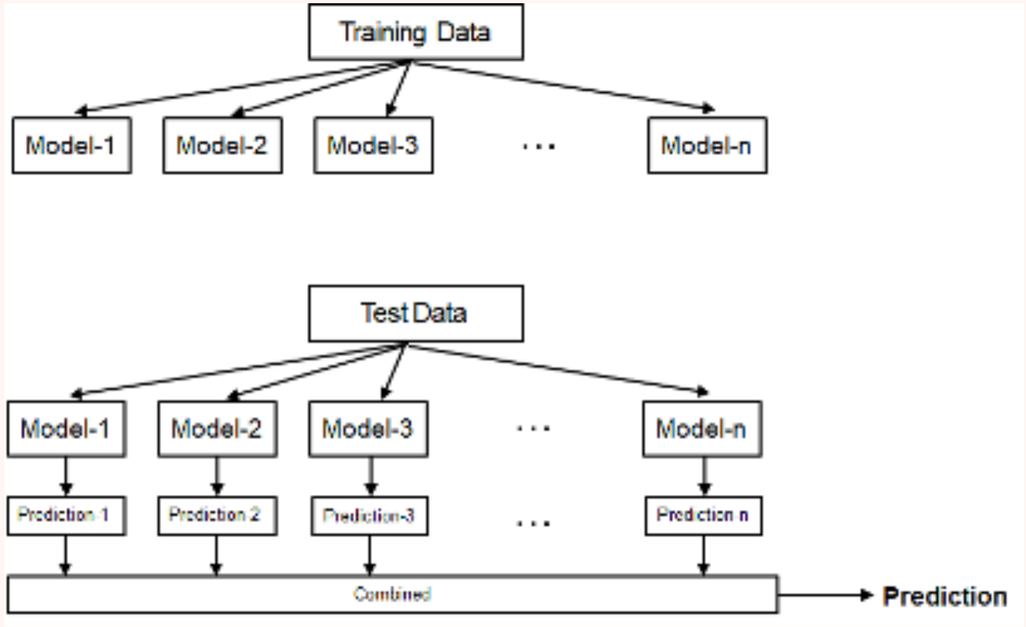
Decision Tree are one of the most simple models

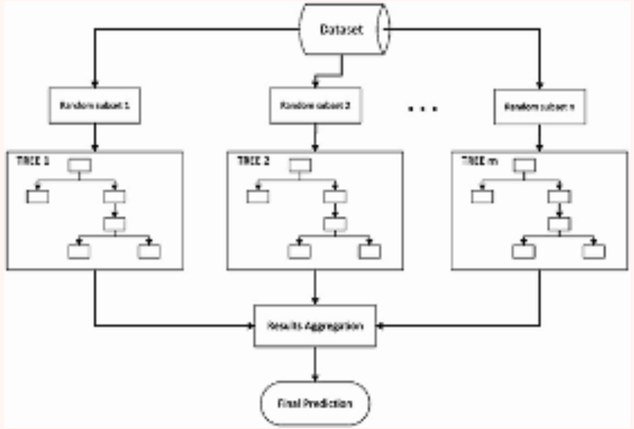


Measure the impurity in a decision tree

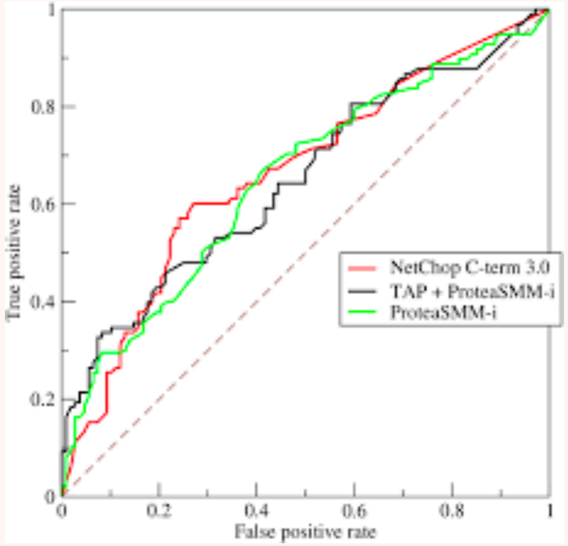
	GINI INDEX	ENTROPY	INFORMATION GAIN	VARIANCE
When to use	Classification Tree	Classification Tree	Classification Tree	Regression Tree
Formula	$G = 1 - \Sigma^{c}_{i=1}(p_i^2)$	$E = -\sum P(X).\log P(X)$	$IG\ (Y, X) = E(Y) - E(Y\|X)$	$V = \Sigma(x-\mu)^2/N$
Range	0 to 0.5 0 = most pure 0.5 = most impure	0 to 1 0 = most pure 1 = most impure	0 to 1 0 = less gain 1 = more gain	-
Characteristics	Easy to compute Non-additive	Computationally intensive Additive	Computationally intensive	The most common measure of dispersion

Ensemble Learning (Bagging and Boosting Techniques)





Accuracy measurement and common problems faced



AIC/BIC values, F1 score,
Matthew Relation Coefficients, KS

Overfitting:reduce the complexity of the model
Imbalance dataset: over/under sampling, SMOTE
NULL values: Mean, median, mode, knn imputation
Outliers present: set the top and floor of the values
Categorical variables: one hot encoding

Time to hit the code!

QUESTIONS PLEASE

THANKS!