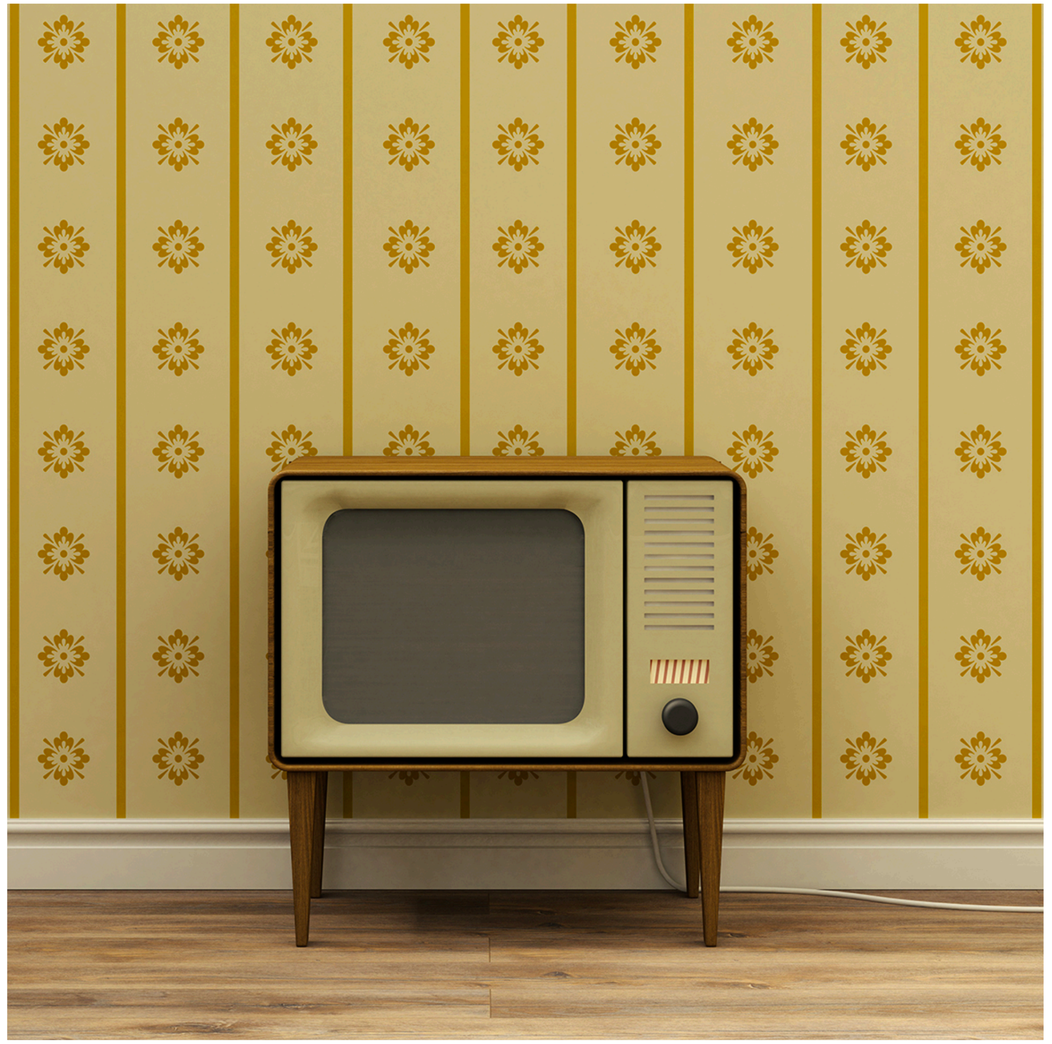
MACHINE LEARNING SESSION 2

Vaibhav Verdhan

Jan 15, 2021

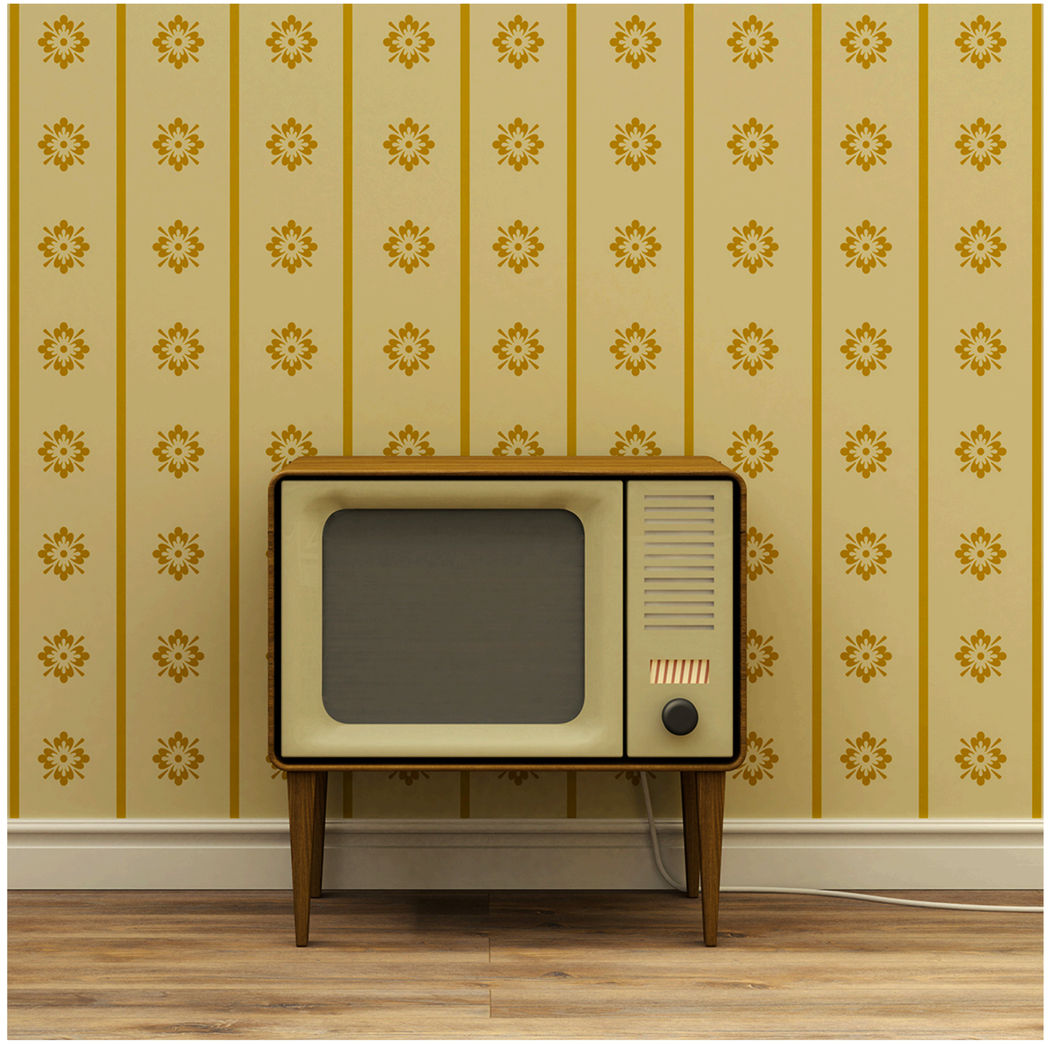
AGENDA FOR THE SESSIONS

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
- Regression Analysis
- Use Cases of Regression
- Simple Linear Regression
- Multiple Linear Regression
- Decision Tree
- Random Forest
- Python Implementations



Regression is used to measure the relationship

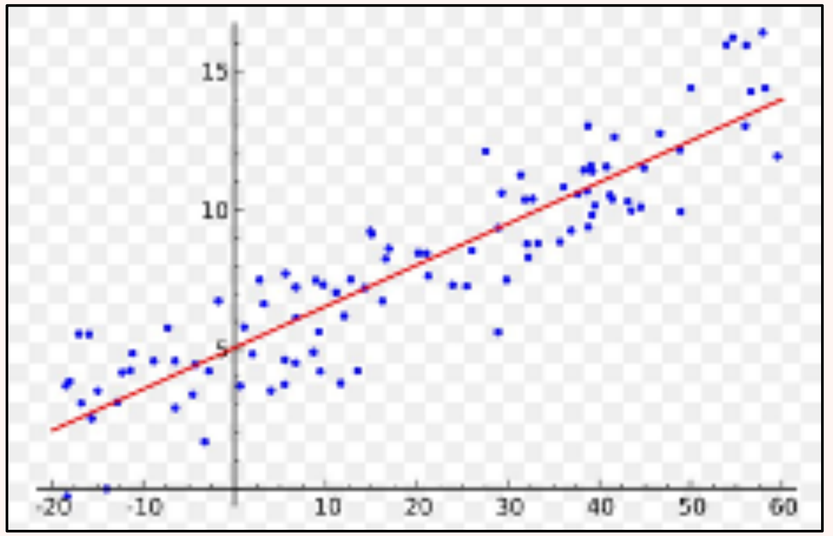
- Linear regression is a way to identify a relationship between the independent variable(s) and dependent variable

- We can use these relationships to predict values for one variable for given value(s) of other variable(s)

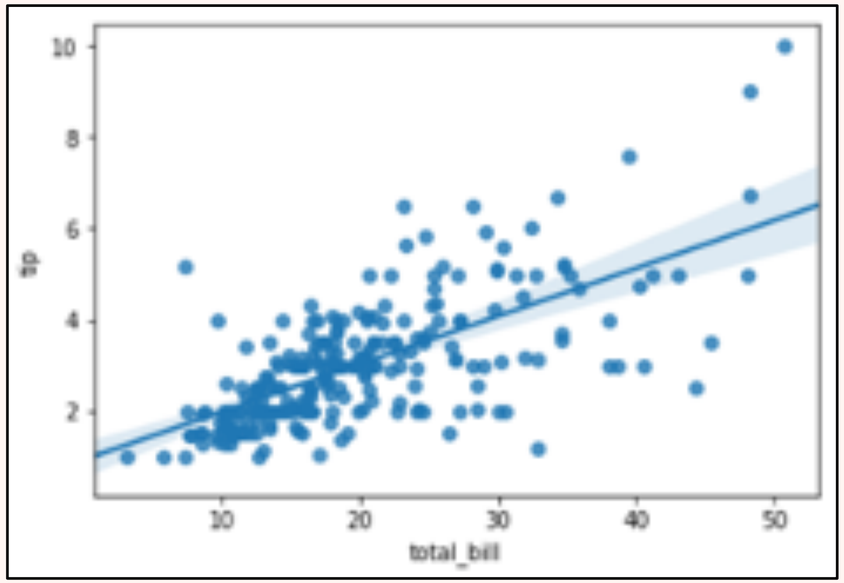
- It assumes the relationship between variables can be modeled through linear equation or an equation of line.

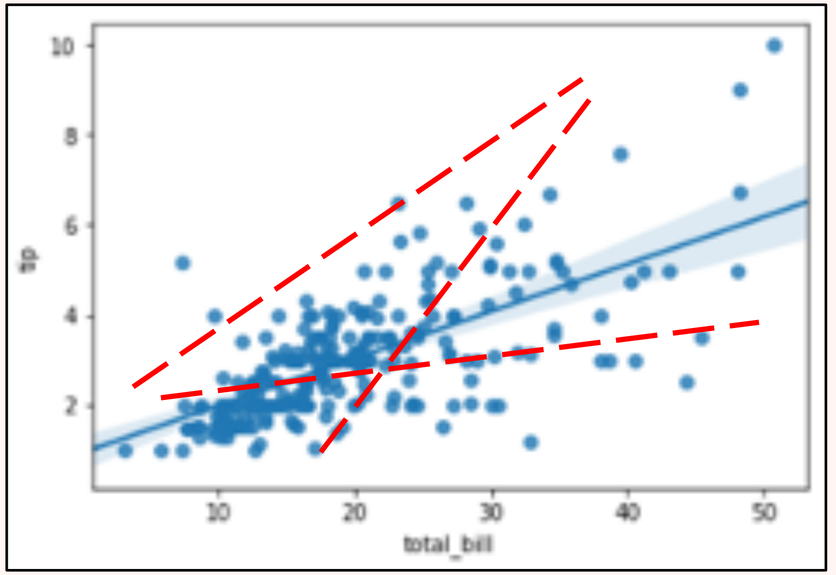
- The variable, which is used in prediction is termed as independent/ explanatory/regressor where the predicted variable is termed as dependent/target/response variable.

- In case of linear regression with a single explanatory variable, the linear combination can be expressed as :
response = intercept + constant * explanatory variable



The objective is to find the line of best fit having min error





$$Y = mx + c$$

$$Y = a*\text{first variable} + b*\text{second variable} + \ldots c$$

Measure the performance of the regression model

R-squared	Adjusted R-squared	Mean Absolute Error	Root Mean Square Error
• Measure of the % of variance in the target variable explained by the model • Generally the first metric to look at for linear regression model performance • Higher the better	• Conceptually, very similar to R-squared but penalizes for addition of too many variables • Generally used when you have too many variables as adding more variables always increases R^2 but not Adjusted R^2 • Higher the better	• Simplest metric to check prediction accuracy • Same unit as dependent variable • Not sensitive to outliers i.e. errors doesn't increase too much if there are outliers • Difficult to optimize from mathematical point of view (pure maths logic) • Lower the better	• Another metric to measure the accuracy of prediction • Same unit as dependent variable • Sensitive to outliers - errors will be magnified due to square function • But has other mathematical advantages that will be covered later

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

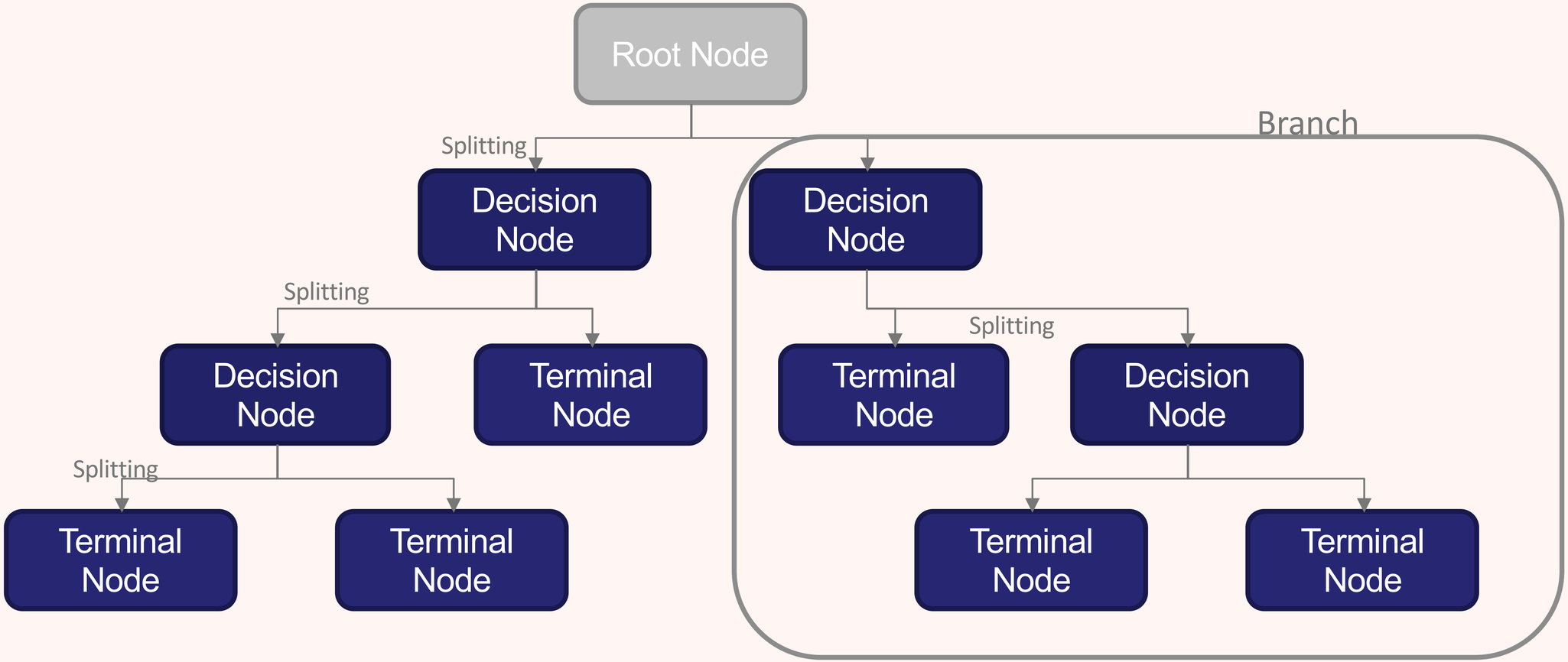
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - yhat_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}$$

Assumptions in a Linear Regression model

Assumption	How to test	How to fix
There should be a linear relationship between dependent and independent variables	Pairplot / Correlation of each independent variables with dependent variable	Transform variables that appear non-linear (log, square root etc.)
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance inflation factor)	Remove correlated variables or merge them
No Heteroskedasticity - residuals should have constant variance	Plot residuals vs. fitted values and check the plot	Non-linear transformation of dependent variable or add other important variables
Residuals must be normally distributed	Plot residuals or use Q-Q plot	Non-linear transformation of independent or dependent variable

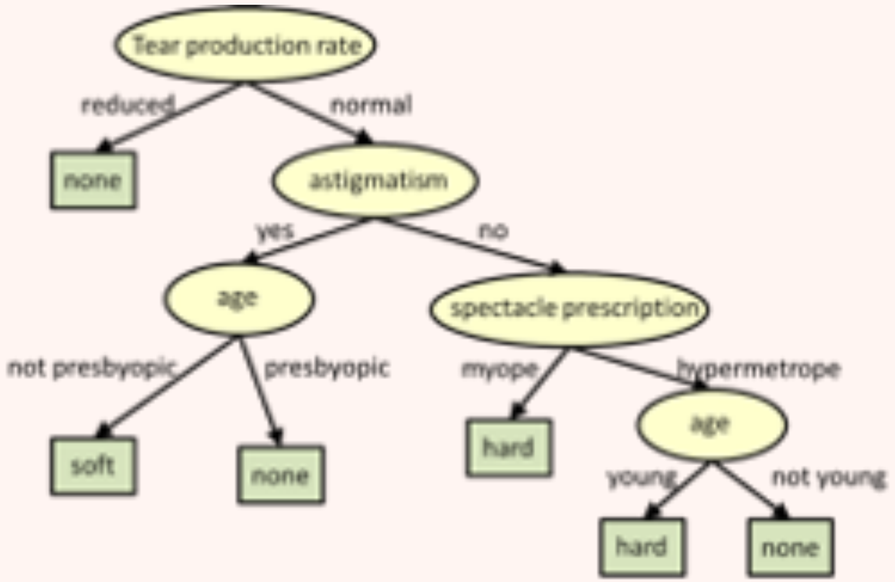
Decision Tree are one of the most simple models

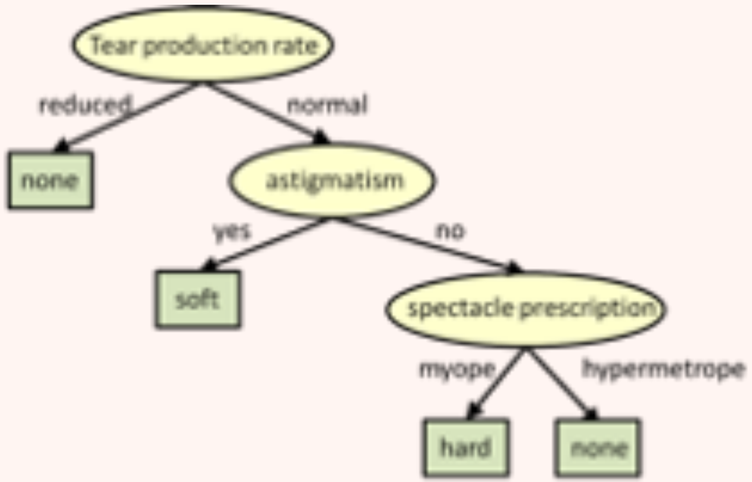


Measure the impurity in a decision tree

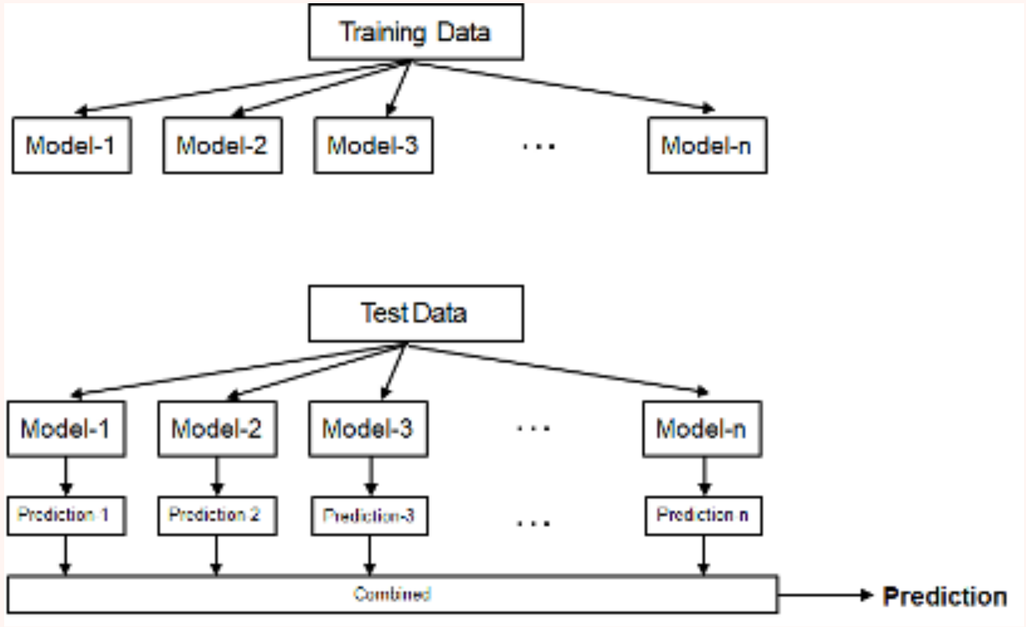
	GINI INDEX	ENTROPY	INFORMATION GAIN	VARIANCE
When to use	Classification Tree	Classification Tree	Classification Tree	Regression Tree
Formula	$G = 1 - \Sigma^{c}_{i=1}(p_i^2)$	$E = -\sum P(X).\log P(X)$	$IG\ (Y, X) = E(Y) - E(Y\|X)$	$V = \Sigma(x-\mu)^2/N$
Range	0 to 0.5 0 = most pure 0.5 = most impure	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

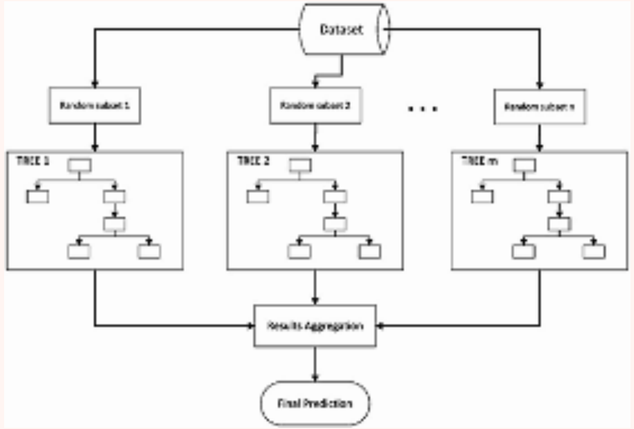
Tackle overfitting in a decision tree





Ensemble Learning (Bagging and Boosting Techniques)





Time to hit the code!

QUESTIONS PLEASE

THANKS!