Generated Notes

* **Random Forest: Predicting Tennis with Data!**
* Random Forest is a powerful computer brain (algorithm) that uses many decision trees to make predictions.
* **What's a Decision Tree?**
* Imagine a "choose your own adventure" book.
* Decision Trees ask yes/no questions to figure things out.
* *Example:* Predicting who survived the Titanic disaster.
* *Question 1:* Did they pay a lot for their ticket?
* *Question 2:* Were they in first class?
* If yes to both, the tree predicts they survived!
* **Building a Tree (Super Simple!)**
* Start with all the data (like Titanic passenger info).
* Find the question that best separates survivors from non-survivors (like ticket class).
* Split the data based on the answer.
* Keep asking questions and splitting until each group is "pure" (all survivors or all non-survivors).
* **Tennis Data Time!**
* Gather tons of tennis match details (wins, losses, player stats).
* Clean the data (fix errors, remove blanks).
* Calculate important stats like:
* **Head-to-Head:** How often players have won/lost against each other.
* **Age/Height Difference:** How old/tall players are compared to their opponents.
* **ELO Rating:** A score showing how good a player is.
* **ELO Rating: Like a Tennis Scoreboard**
* ELO shows a player's skill level (used in chess too!).
* If a player wins, their ELO goes up. If they lose, it goes down.
* *Example:* Roger Federer's ELO went up as he won more matches.
* Can be specific to different surfaces (clay, grass, hard). Nadal is the "King of Clay" due to his high clay ELO.
* **Decision Tree vs. Random Forest**
* **Decision Tree:**
* Simple, but can be shaky (sensitive to the data).
* **Random Forest:**
* Many Decision Trees working together.
* Each tree uses a random sample of the data.

* The forest votes on the final prediction.
* More stable and accurate than a single tree.
* **XG Boost: Random Forest on Steroids!**
* Like a super-powered Random Forest.
* Prevents trees from growing too big.
* Even more accurate!
* **Testing the Models:**
* Used tennis data up to December 2024 to train the models.
* Then predicted the winners of the 2024 Australian Open (which the models hadn't seen).
* XG Boost model correctly predicted 99 out of 116 matches!
* **Key Takeaways**
* Random Forests and XG Boost are great for predicting outcomes using data.
* ELO rating is a useful way to measure a player's skill.
* Using many trees (Random Forest) is better than using just one (Decision Tree).