SLIDE 1:
Hello! Today I am going to be presenting my study on “Impact of Family background and Socioeconomic Status on Marijuana Use”. My models are decision trees, the simplest way to interpret data.

SLIDE 2:
Read through the slide interactively.

SLIDE 3:
Explain decision trees-definition, optimisation.
Same for bagging, boosting, random Forests.

SLIDE 5:
Explain the project stepwise. How you arrived at each step of it. Start by data cleaning, exploring, modeling.

SLIDE 6&7&8:
All the necessary content is on the slides.

SLIDE 9:
In this assignment, I worked with three types of variables: IMOTHER (binary), INCOME (multi-class categorical), and IRMJFY (numerical). Binary variables, like IMOTHER, are particularly well-suited for decision trees because they align naturally with the tree’s split structure. Trees tend to favor binary variables, as they offer clear yes/no decisions that reduce complexity. On the other hand, multi-class categorical variables like INCOME can be more challenging—especially if they are ordered categories. If not handled carefully, trees might interpret these classes incorrectly or miss out on the natural order of the data. However, these variables can be powerful when specific class distinctions matter. Numerical variables, such as IRMJFY, are useful for identifying threshold-based splits, which can reveal important cutoff points in the data.

SLIDE 10:

The importance of variables varies across models, but in the tree shown BEFORE, the most influential features were those related to family structure and emotional support, such as IFATHER and PRPROUD2. This makes intuitive sense—parental presence and pride can strongly influence a child's behavior.

As responsible data science communicators, it's essential to scale or anonymize sensitive data to protect individual privacy—especially when dealing with topics like family income or government assistance.

Also, we must avoid overgeneralizing model outputs. Just because a child doesn’t receive allowance for chores doesn’t mean they are more likely to use marijuana. These are statistical patterns, not personal truths.

Finally, always approach sensitive datasets with care and neutrality. The goal is insight, not judgment.

SLIDE 11&12:

READ THROUGH THEM MOSTLY//ADD POINTS WHENEVER NECESSARY.