

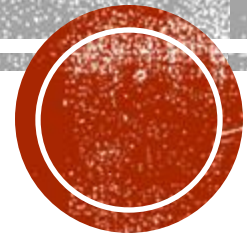
# PREDICTING LEGISLATIVE OUTCOMES: VOTER PREDICTIONS IN CONGRESS

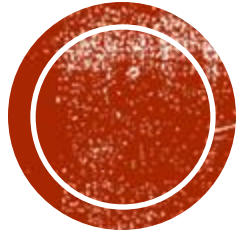
## A CRISP-DM Approach to Predictive Modeling

**Project Domain:** Politics

**Project Team:** Summer Survivors

**Contributors:** Raul Garcia, Daniel McCarty, Kim Sundblom, Mason Wester, and Paige Winnett





# BUSINESS UNDERSTANDING

- What is the problem you're trying to solve?
- Identifying stakeholders
- Understanding the motivation to find answers
- Articulating the action plan



# BUSINESS UNDERSTANDING

- Business Problems (What)

➔ Provide a way to predict voter outcomes to anticipate the passing of a bill in Congress.

The sample objective is to predict whether an "education-spending" bill introduced in Congress will receive the necessary 172 Yes votes to pass. This will be done by analyzing historical voting patterns and party affiliation of Congressmen, using various classification models.

The specific question to be answered is:

*Will the bill pass based on the **predicted** votes of Congressmen who have not yet voted?*



# BUSINESS UNDERSTANDING

- Stakeholders (Who)

➔ Campaign Managers, Political analysts, Journalists, Policymakers

**This set of stakeholders rely on early insights to make informed decisions. Policymakers can understand voting patterns to make data-driven legislative decisions. Campaign managers, for example, could use these predictions to fine-tune their outreach strategies, while journalists could publish early projections about the bill's fate.**

The core stakeholder mix includes:

*Those directly involved in the outcomes of political issues,  
and/or those within the ecosystem that govern them.*



# BUSINESS UNDERSTANDING

- Motivation (Why)



The added value comes from predicting voter outcomes to provide strategic advantage and impact political strategy

The ability to predict the outcome of a congressional vote before it happens is crucial for strategic decision-making in politics. It can provide valuable insights into the likelihood of passing important legislation like in our example, the education-spending bill. By accurately predicting the outcome, stakeholders can take early action, whether to rally support, adjust messaging, make policy decisions, or communicate to the masses.

Key consideration as to the motive:

*Predicting vote outcomes provides strategic advantages, helping to influence decisions before a result.*



# BUSINESS UNDERSTANDING

- Action Planning (How)

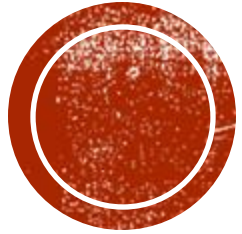
- ➔ Train a classification model using historical vote data and party affiliation
- ➔ Apply the model to a new dataset to predict the voting outcomes
- ➔ Use the predictions to determine if the bill will secure the necessary votes

Using historical voting data, we will train a classification model that predicts Yes, No, or Abstain votes based on party affiliation and past voting behavior. The model will then be applied to a new dataset, where the voting outcome for the education-spending bill is missing, to predict how each Congressman will vote. This will help determine whether the bill will meet the threshold of the required 172 Yes votes.

Core objectives of the action plan:

*Developing the model will solve our business problem by forecasting the likelihood of a bill passing, enabling data driven decisions by stakeholders.*





# DATA UNDERSTANDING

- How many datasets do you have?
- Record and Attribute count
- Understanding each attribute
- Variable types
- Target attribute



## Using Two Seperate provided .CSV

	Vote-Train.csv	Vote-Predict-Vote.csv
Attributes	17	17
Records	400	35



Each attribute is a polynomial offering 3 options in how each person voted on different topics:

Yes(y)

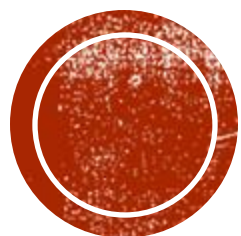
No(n)

Abstain

The target attribute was education spending; to use the vote-train.csv to predict how the voters in vote-predict-vote.csv would vote on education spending







# DATA PREPARATION

- Any missing values, outliers, or unstructured data?
- Any highly-correlated attributes?
- Redundant attributes?
- Irrelevant attributes?
- Other data quality issues?





- **Are there any missing values, outliers, or unstructured data?**

Understanding that missing values can lead to biased results or inaccurate modeling predictions is vital to data preparation. While our current data set did not have any missing values or or unstructured data, it is important to always check the data sets to know whether to add or remove incomplete entries to ensure data integrity.

- **Any highly-correlated attributes?**

It's important to check for highly-correlated attributes because when features are too similar, they provide redundant information, which can confuse the model and reduce its accuracy. This can lead to issues like multicollinearity, where the model struggles to determine which attribute is driving the outcome, potentially resulting in unstable or less interpretable results. By identifying and removing highly-correlated features, we ensure the model is more efficient, easier to understand, and better at making reliable predictions.

- **Redundant attributes?**

Redundant attributes create duplicate information, adding extra complexity without contributing value to the analysis. It is important to remove any overlapping data to keep things simple to improve efficiency. Our data sets did not have any redundant attributes.





- **Irrelevant attributes?**

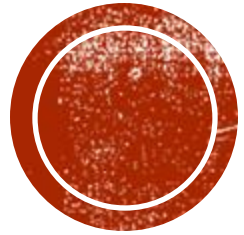
Even though our data sets did not have irrelevant information due to all voting data being used in the model, it is still extremely important to remove the irrelevant data from the data sets to maintain accuracy of the model. Irrelevant attributes create unnecessary information that will reduce the accuracy of the model with the more irrelevant attributes there are.

- **Other data quality issues?**

The voting data sets used did not have any other data quality issues. The data was double checked for inconsistent data forms, duplicate entries, and incorrectly recorded data. These can create errors in the analysis leading to inaccurate results.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	handicap	water-prc	adoption	physician	el-salvado	religious-	anti-satell	aid-to-ru	ma-missi	immigrat	synfuels-	educator	superfund	crime	duty-free	export-ad	Party
1	n	y	n	y	y	y	n	n	n	n	abstain	y	y	y	n	y	republican
2	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n	abstain	republican
3	abstain	y	y	abstain	y	y	n	n	n	n	y	n	y	y	n	n	democrat
4	n	y	y	n	abstain	y	n	n	n	n	y	n	y	n	n	y	democrat
5	y	y	y	n	y	y	n	n	n	n	y	abstain	y	y	y	y	democrat
6	n	y	y	n	y	y	n	n	n	n	n	n	y	y	y	y	democrat
7	n	y	n	y	y	y	n	n	n	n	n	n	y	y	y	y	democrat
8	n	y	n	y	y	y	n	n	n	n	n	n	abstain	y	y	y	democrat
9	n	y	n	y	y	y	n	n	n	n	n	n	y	y	abstain	y	republican
10	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n	y	republican
11	y	y	y	n	n	n	y	y	y	n	n	n	n	n	abstain	abstain	democrat
12	n	y	n	y	y	n	n	n	n	n	abstain	abstain	y	y	n	n	republican
13	n	y	n	y	y	y	n	n	n	n	y	abstain	y	y	abstain	abstain	republican
14	n	y	y	n	n	n	y	y	y	n	n	n	y	n	abstain	abstain	democrat
15	y	y	y	n	n	y	y	y	abstain	y	y	abstain	n	n	y	abstain	democrat
16	n	y	n	y	y	y	n	n	n	n	n	y	abstain	abstain	n	abstain	republican
17	n	y	n	y	y	y	n	n	n	y	n	y	y	abstain	n	abstain	republican
18	n	n	y	n	n	y	n	y	abstain	y	y	y	abstain	n	n	y	democrat
19	y	abstain	y	n	n	n	y	y	y	n	n	n	y	n	y	y	democrat
20	n	y	n	y	y	y	n	n	n	n	n	abstain	y	y	n	n	republican
21	y	y	y	n	n	n	y	y	y	n	y	n	n	n	y	y	democrat
22	y	y	y	n	n	abstain	y	y	n	n	y	n	n	n	n	y	democrat
23	y	y	y	n	n	n	y	y	y	n	n	n	abstain	abstain	y	y	democrat
24	y	abstain	y	n	n	n	y	y	y	n	n	abstain	n	n	y	y	democrat
25	y	y	y	n	n	n	y	y	y	n	n	n	n	n	y	y	democrat
26	n	n	y	n	n	n	y	y	y	n	n	n	n	n	y	y	democrat
27	y	n	y	n	n	n	y	y	y	n	y	n	n	n	y	y	democrat
28	y	y	y	n	n	n	y	y	y	n	y	n	n	n	y	y	republican
29	y	n	n	y	y	n	y	y	y	n	n	y	y	y	n	y	republican
30	y	y	y	n	n	n	y	y	y	n	y	n	n	n	y	y	democrat
31	y	y	y	n	n	n	y	y	y	n	n	n	y	y	n	y	democrat
32	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n	n	republican





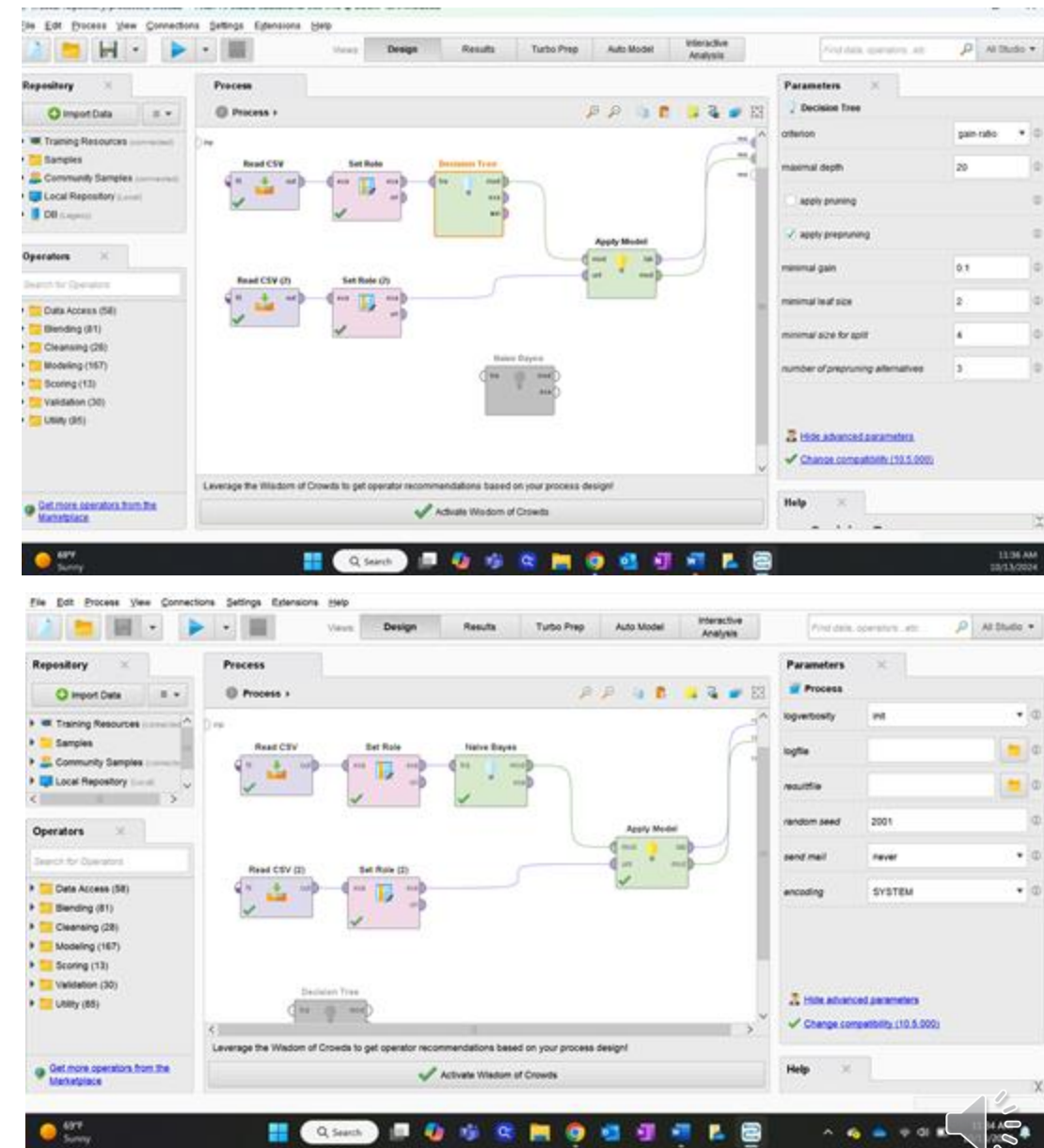
# MODELING



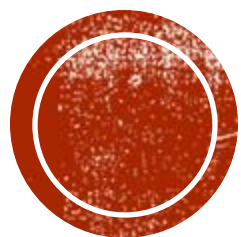
# MODELING

## Models Used:

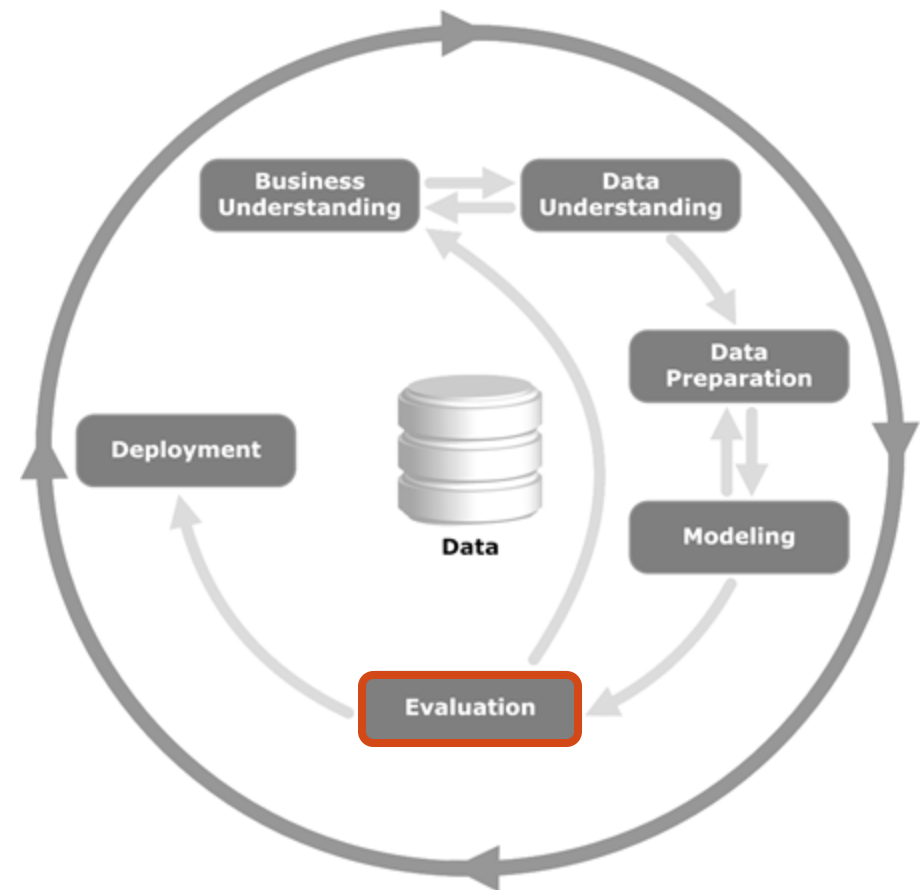
- **Decision Tree:** Classifies votes by creating decision rules based on voting patterns. Built using gain ratio, maximal depth of 20, uncheck pruning, check prepruning, with a minimal gain of 0.1, minimal leaf size of 2, minimal size of split of 4, # prepruning alternatives of 3 (followed wk5 lab)
  - **What it provided:** It showed clear paths where past votes helped predict Educational Spending outcomes, offering insights into which votes had the most impact. The Decision Tree provides a model that makes it easier to explain which past decisions influenced voting behavior, but it had slightly lower accuracy compared to Naive Bayes.
- **Naive Bayes:** Uses probability distributions to predict outcomes based on past votes. Trained with Laplace correction to handle zero-probability issues in categorical data.
  - **What it provided:** The Naive Bayes model produced higher accuracy than Decision Tree and was able to generalize better across unseen data. Although less interpretable than Decision Tree, Naive Bayes excels in prediction accuracy, making it the best choice for predicting future votes.
- **Neural Network:** This model was not used because the data is predominantly categorical, and Neural Networks work better with numerical data. Instead, we focused on models suitable for categorical inputs.



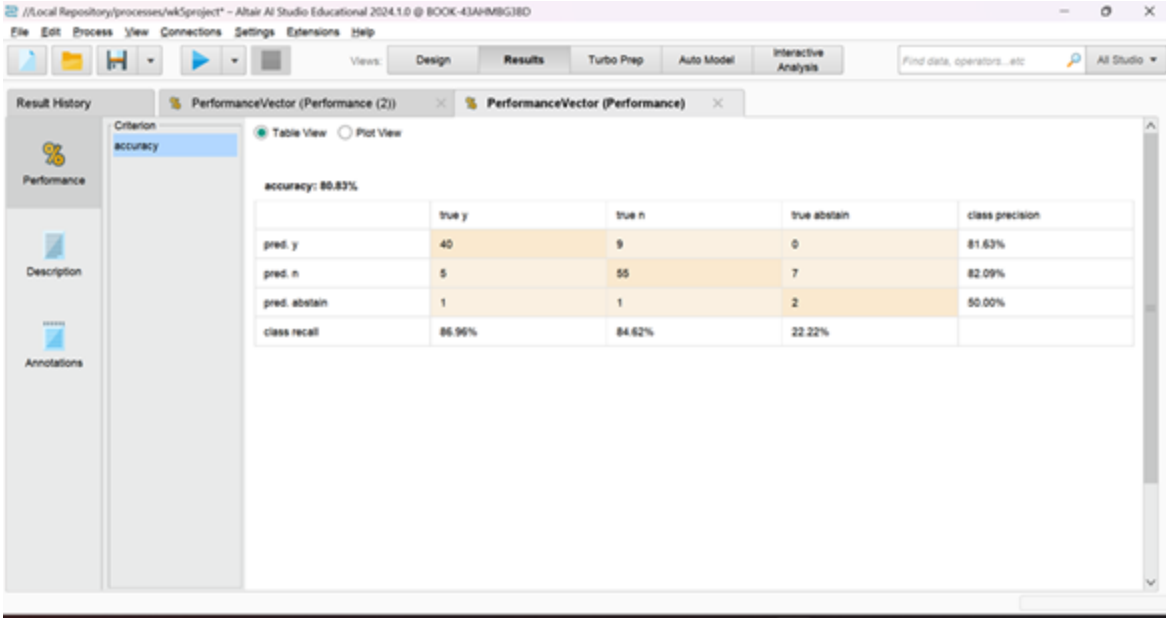
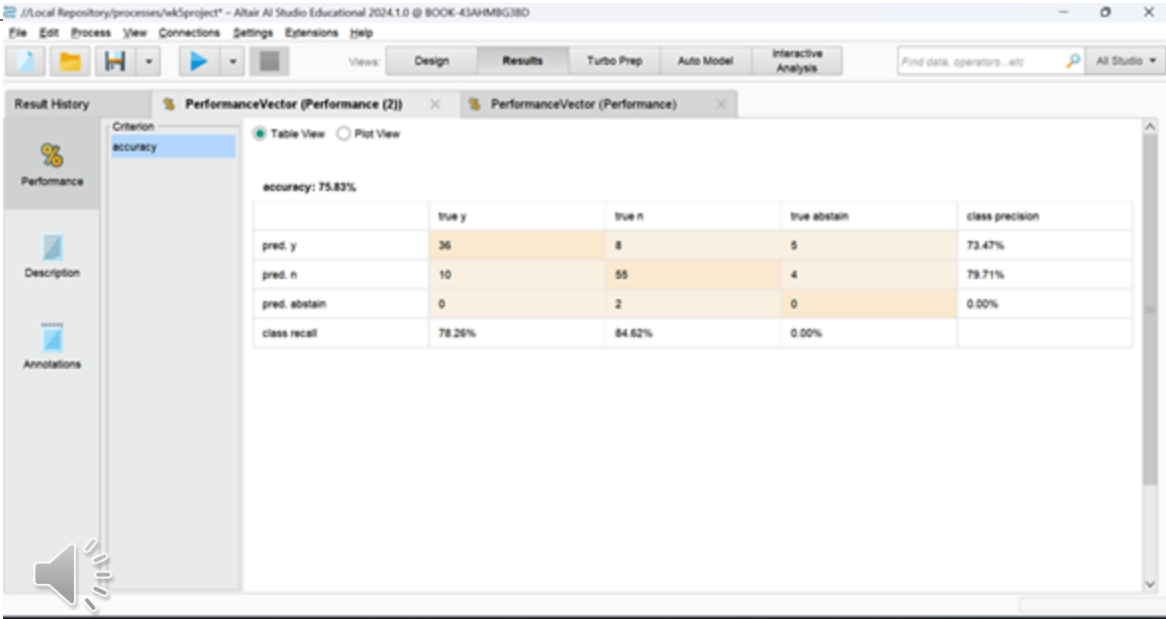
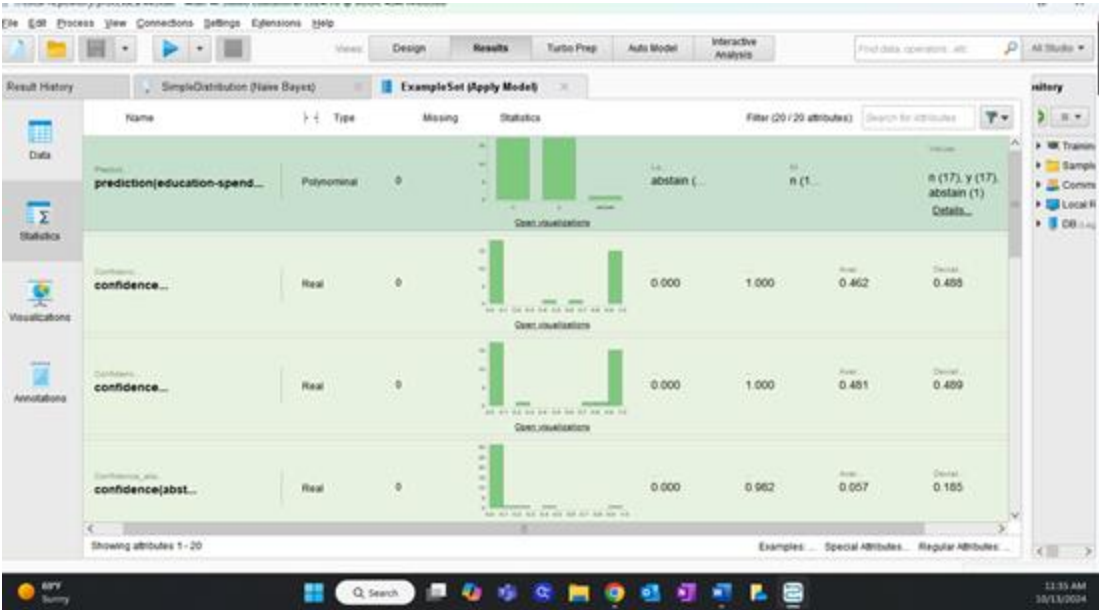
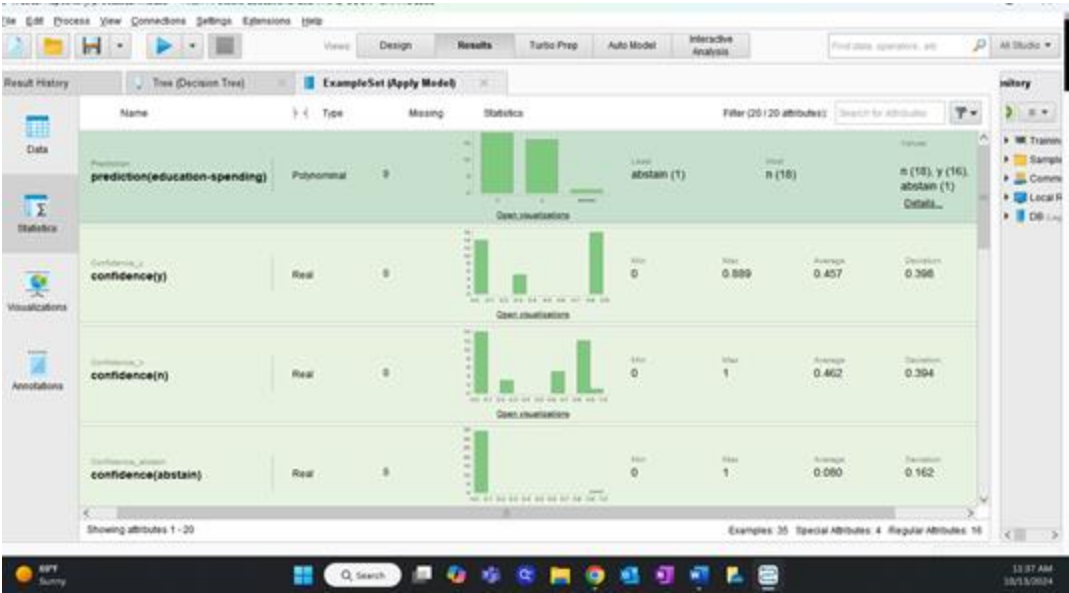




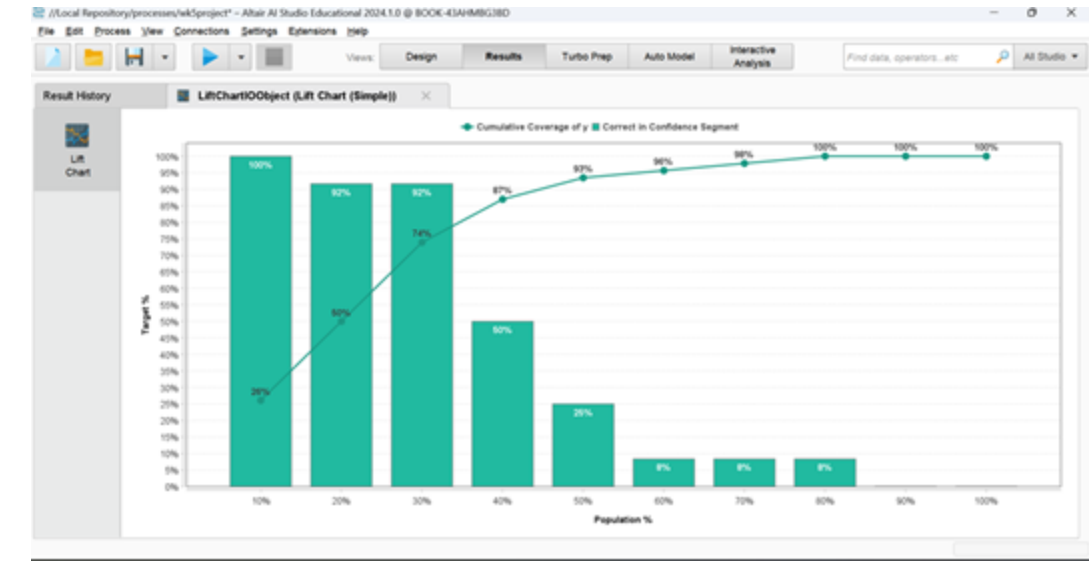
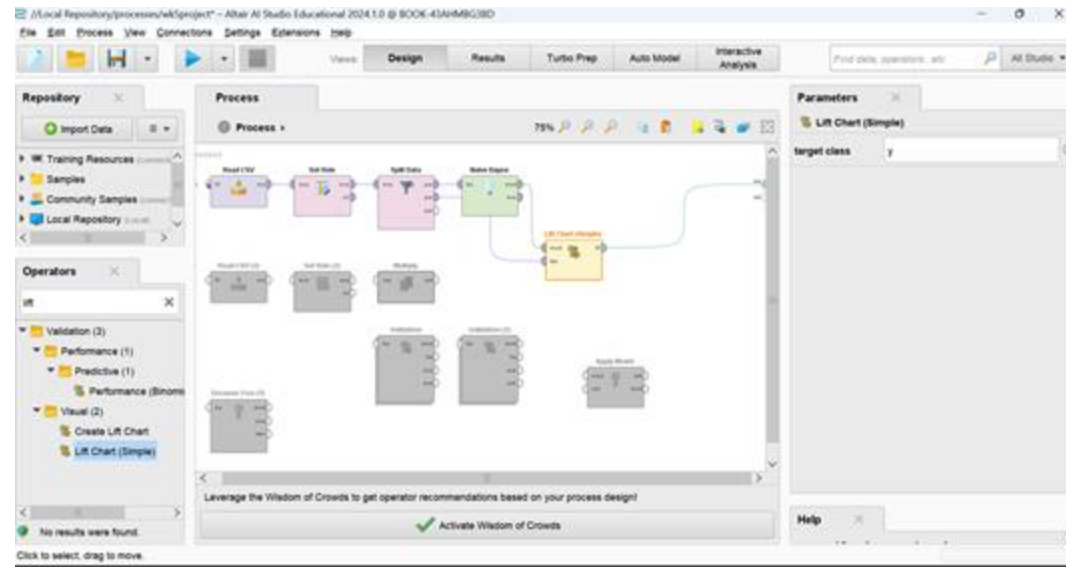
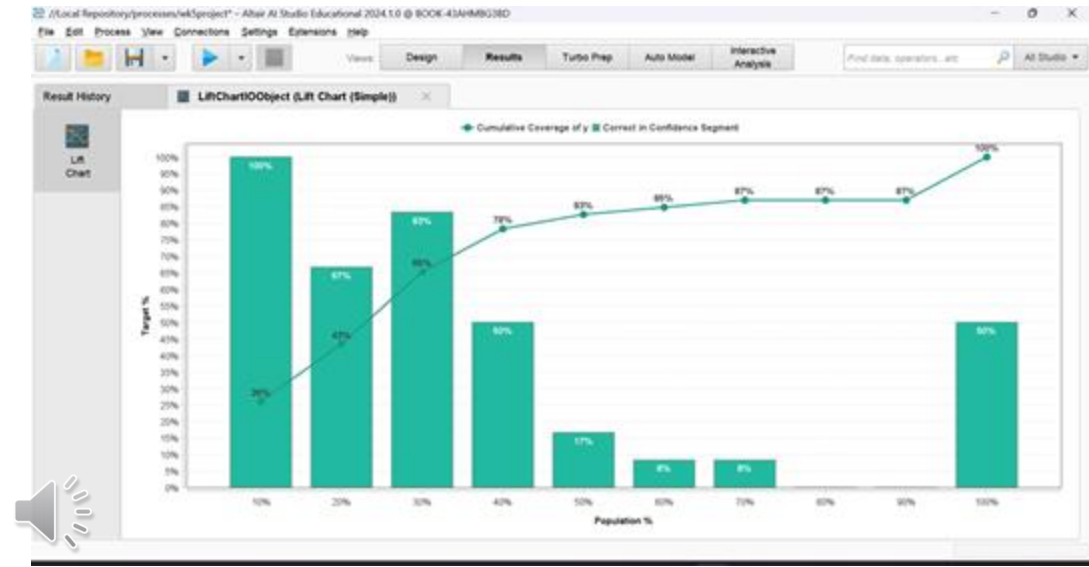
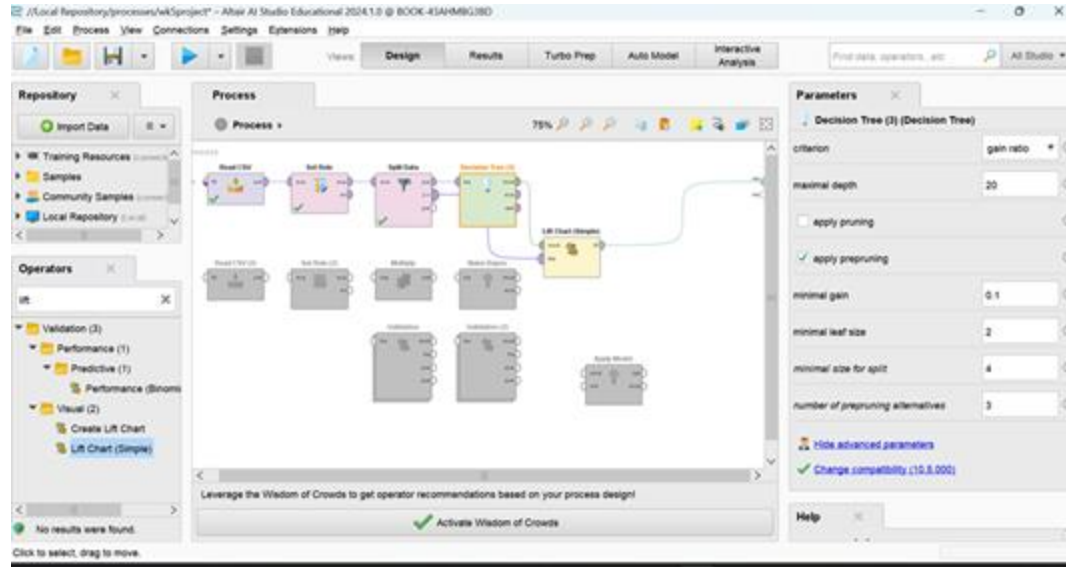
# EVALUATION



# Evaluation



# Evaluation






# Evaluation

---

## Decision Tree:

- **Confidence:** 88.9% for "Yes" votes, predicting 16 "Yes" votes.
- **Precision:** 73.47% for predicting "Yes" votes.
- **Accuracy:** 75.83%.
- **Recall:** 78.26%, meaning the model identified most of the true "Yes" votes.
- **Lift Chart:** Demonstrated strong confidence in identifying top-ranked "Yes" votes with 100% confidence in the top 10% of the population with cumulative coverage of 26%.

## Naive Bayes:

- **Simple Distribution:** Yes = 0.388, No = 0.540, Abstain = 0.072 
- **Confidence:** 100% for "Yes" votes, predicting 17 "Yes" votes.
- **Precision:** 81.63% for predicting "Yes" votes.
- **Accuracy:** 80.83%, outperforming Decision Tree.
- **Recall:** 86.96%, identifying a higher proportion of true "Yes" votes.
- **Lift Chart:** Similar to Decision Tree, Naive Bayes also provided 100% confidence in top 10% of predictions and cumulative coverage of 26%, but with significantly better confidence and cumulative coverage in top 20% (when compared to top 20% for the decision tree).

## Comparison and Selection:

- **Naive Bayes** performed better overall with higher accuracy and recall, making it the preferred model.
- **Naive Bayes** was selected due to its superior accuracy and recall in predicting "Yes" votes, making it the most reliable for the task. The Lift Chart results reinforced its ability to rank important predictions more effectively although both models had strong Lift Chart results
- The **Decision Tree** also performed well but had slightly lower accuracy.



## **Confusion Matrix:**

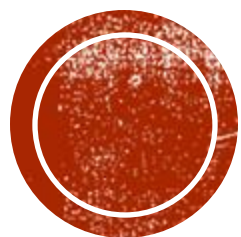
- The confusion matrix for both the Decision Tree and Naive Bayes models shows that Naive Bayes had a higher number of true positives, meaning it was better at predicting "Yes" votes. This aligns with why we selected it as the final model.

## **Lift Chart:**

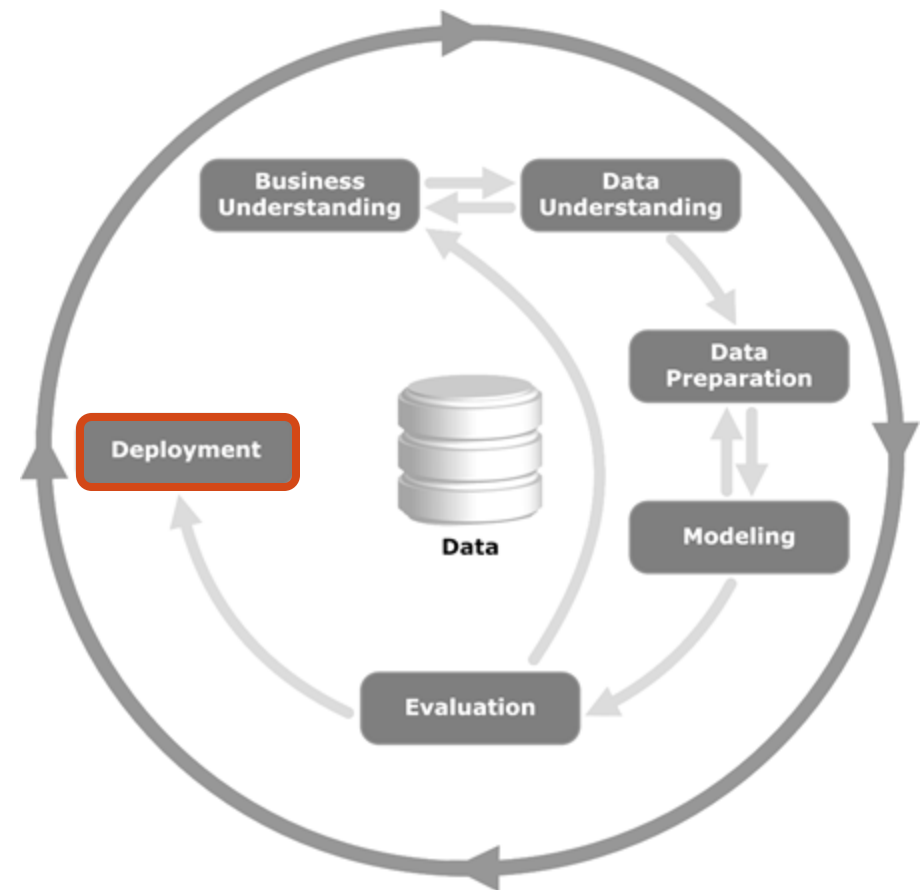


- In our case, the Naive Bayes model had 100% confidence in predicting "Yes" votes in the top 10% of the data, meaning every prediction for "Yes" in this top tier was correct. This is a crucial finding because it shows that the model is very good at prioritizing the most likely supporters of the bill, which is essential for stakeholders who need to focus their resources on securing votes.
- Both the confusion matrix and lift chart demonstrate that Naive Bayes was more effective than the Decision Tree in identifying the critical "Yes" votes, making it the better model for our task.





# DEPLOYMENT



# DEPLOYMENT

---

## Value of the Project:

- The Naive Bayes model provides accurate predictions for Congressmen's voting decisions on Educational Spending. This enables stakeholders such as political analysts, campaign strategists, and policymakers to anticipate voting outcomes and plan accordingly.

## How the Model Can Be Used:

- The model can be integrated into a **decision support system** where users input past voting records, and the system predicts future voting behavior on similar bills.
- This predictive capability allows campaign teams to tailor their messaging or advocacy efforts based on likely voting patterns, thereby increasing the effectiveness of lobbying or public relations campaigns.

## Who Will Benefit:

- **Campaign Managers & “Business” Lobbyists:** Can use the predictions to refine strategies for influencing voting outcomes.
- **Policymakers & Economists:** Can forecast the likelihood of a bill passing and adjust their legislative approaches.
- **Journalists:** Can analyze and report on potential voting outcomes to inform the public.



# DEPLOYMENT

- **Project Value**

**Value (why deploy):**

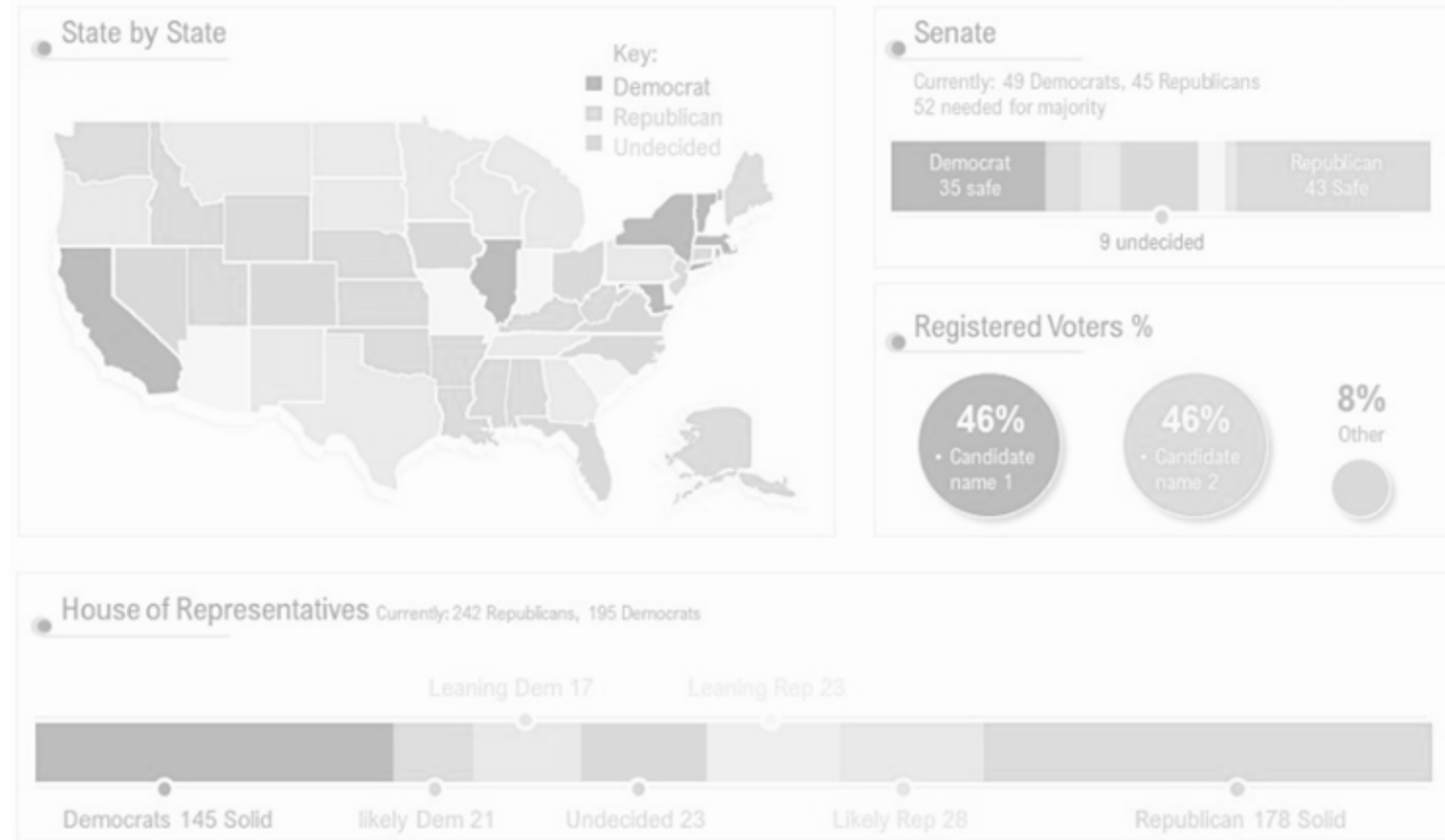
- Provides actionable insights before critical votes, enabling faster, more informed decision-making.

- **Stakeholder Perspective**

**Impact (who benefits):**

- Campaign managers, political analysts, journalists, and policymakers will access real-time predictions to support strategic decision-making.

Dashboard Depicting Political Voting Outcomes

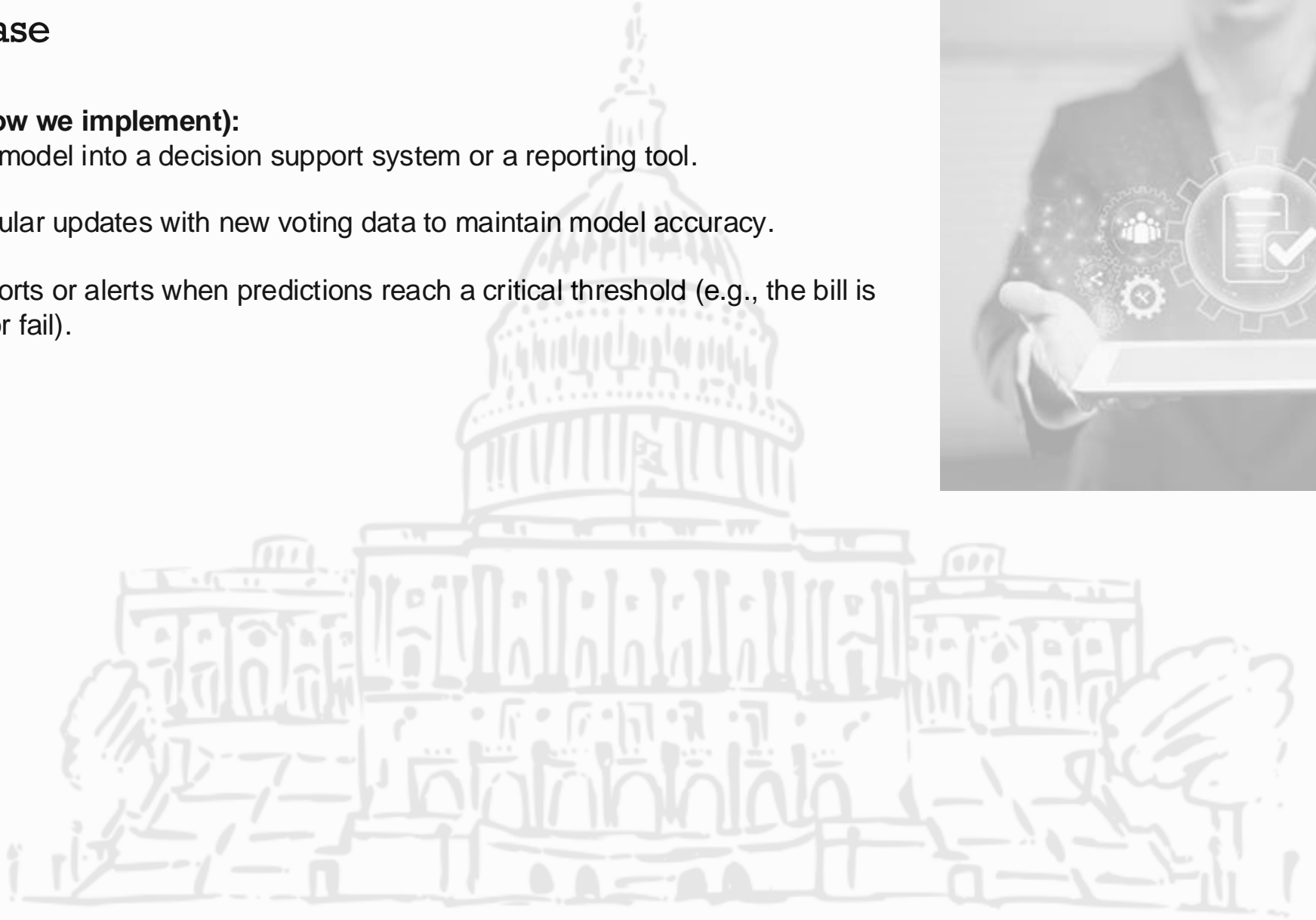


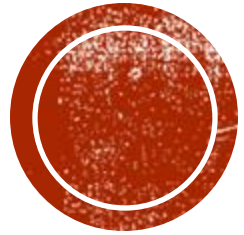
# DEPLOYMENT

- **Model Use Case**

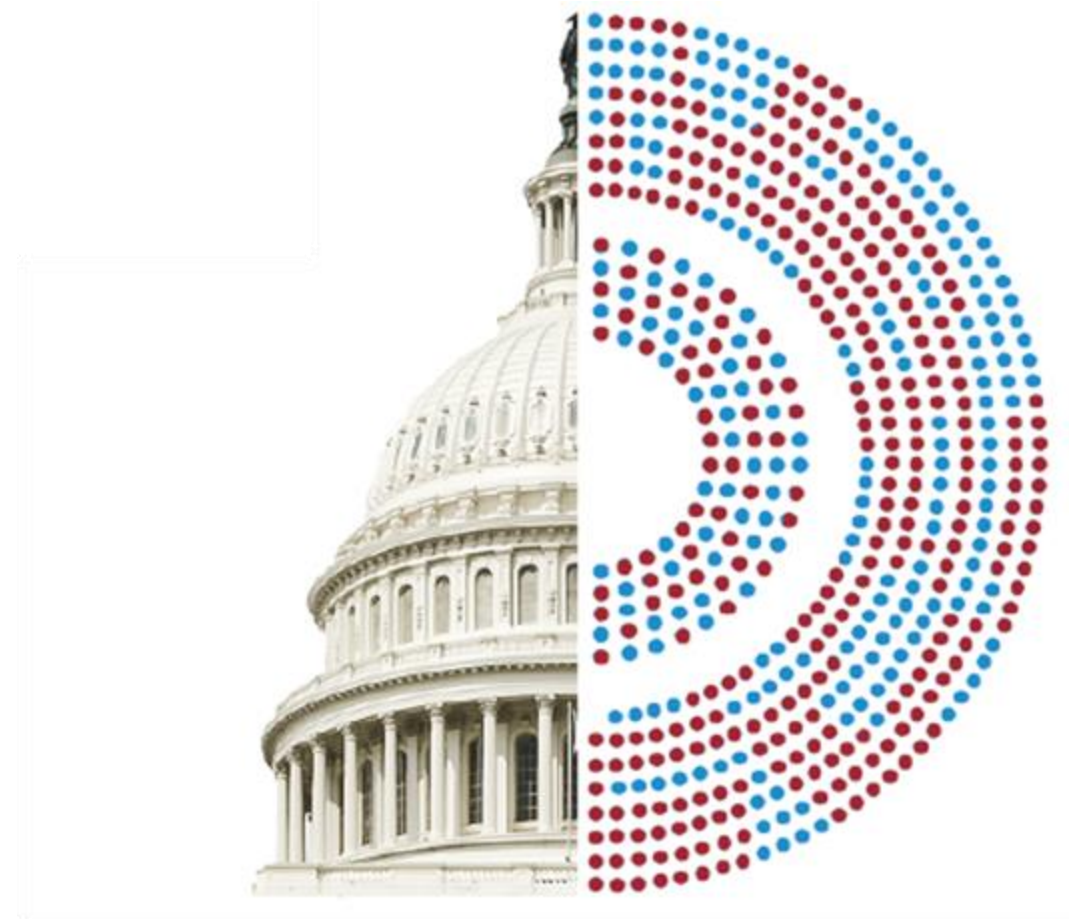
**Execution (How we implement):**

- Integrate the model into a decision support system or a reporting tool.
- Schedule regular updates with new voting data to maintain model accuracy.
- Generate reports or alerts when predictions reach a critical threshold (e.g., the bill is likely to pass or fail).





# REFLECTION



- Choose 4 of the following reflection suggestions on the rubric



# REFLECTION

---

## **The Importance of Data Mining:**

- This project demonstrated the power of data mining to uncover patterns in historical voting data, making it possible to predict future behavior with significant accuracy. The use of both Decision Tree and Naive Bayes models reinforced how different techniques can be applied to real-world political problems.
- Data-mining with proper training data can be used to find patterns and predict possible outcomes in many different situations. Predicting buying habits, life of a machinery, estimating time to complete tasks, and many other important things.

## **Challenges and Solutions:**

- Challenge: Understanding the nuances of the models, particularly in handling categorical data and the exclusion of the Neural Network model.
  - Some limitations include reliance on data quality, if the data you're using isn't very accurate and complete it can leave gaps that can cause estimations or results to vary significantly.
- Solution: After research and testing, we confirmed that Naive Bayes was a better fit for categorical data, and excluding Neural Network was justified, and our data was accurate and complete.





# REFLECTION

---

## **Lessons Learned:**

- We learned that model selection depends heavily on the nature of the data. While Decision Tree offers more interpretability, Naive Bayes excelled in performance due to its probabilistic approach.
- The importance of data quality and thorough evaluation was evident as we found better results after applying additional validation techniques.
- Our biggest struggle as a group was being able to find times we were all available. We had to use communication and clear instruction in order to get everything done and combined.
  - We leveraged various communication tools, as well as collaborated simultaneously to one project file (ppt) and we were creative when it came to working around everyone's' schedules for recording so we could accommodate everyone while providing a thoroughly organized project.

## **Future Suggestions:**

- Future work could involve testing the model on newer data sets as voting patterns evolve over time, ensuring that the model stays relevant.
- Expanding the model to include additional contextual factors (political climate or economic indicators) could improve prediction accuracy.



# REFLECTION

---

Data-mining with proper training data can be used to find patterns and predict possible outcomes in many different situations. Predicting buying habits, life of a machinery, estimating time to complete tasks, and many other important things.

Some limitations include reliance on data quality, if the data you're using isn't very accurate and complete it can leave gaps that can cause estimations or results to vary significantly.

Our biggest struggle as a group was being able to find times we were all available. We had to use communication and clear instruction in order to get everything done and combined. Ultimately, the members of Summer Survivors are pleased with each individual outcomes and through dedication to the project and the team delivered all required elements for a successful and positive project experience.



# REFERENCES

---

- **Wk5 Lab with provided Data Sets (vote-train.csv & vote-predict-vote.csv)**
- **Wk4Lab**
- **Wk6Lab**
- **Group Project Instructions (pdf)**
- **Book (Linoff&Berry)**
  - Linoff, G. S., Berry, M. J. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Germany: Wiley.
- **Model Evaluation in Tan's book**
  - *Classification: Basic Concepts, Decision Trees, and Model Evaluation*. (n.d.). <https://www-users.cse.umn.edu/~kumar001/dmbook/ch4.pdf>
- **Book (North)**
  - North, M. (2018). *Data mining for the masses : with implementations in RapidMiner and R*. Createspace.
- **Naive Bayes Classifier (Assigned Reading)**
  - Bayes, T. (n.d.). *Naïve Bayes Classifier We will start off with a visual intuition, before looking at the math....*  
[https://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect\\_examples.pdf](https://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf)

