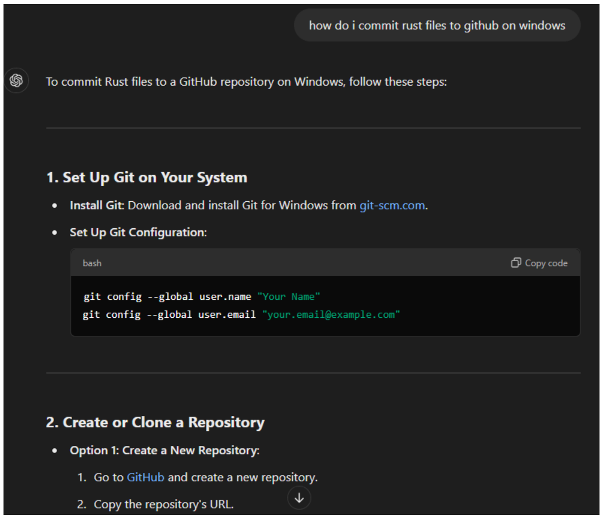
DS 210 Final Project Writeup

Sources used for the final project include lecture notebooks, specifically lectures 21 and 22 for cleaning and decision tree construction, the Rust textbook, the linfa_trees documentation, the ndarray documentation, and ChatGPT to explain how to use Git and commit my changes (Transcript provided below).



how do i commit rust files to github on windows

To commit Rust files to a GitHub repository on Windows, follow these steps:

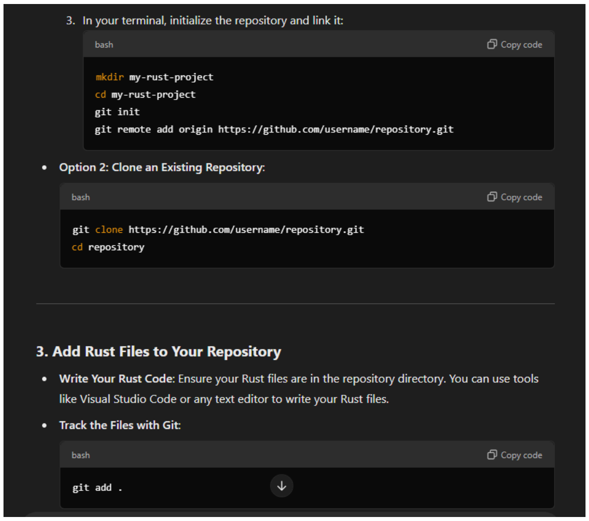
1. Set Up Git on Your System

- **Install Git:** Download and install Git for Windows from git-scm.com.
- **Set Up Git Configuration:**

```bash
git config --global user.name "Your Name"
git config --global user.email "your.email@example.com"
```

2. Create or Clone a Repository

- **Option 1: Create a New Repository:**
  1. Go to GitHub and create a new repository.
  2. Copy the repository's URL.



3. In your terminal, initialize the repository and link it:

```bash
mkdir my-rust-project
cd my-rust-project
git init
git remote add origin https://github.com/username/repository.git
```

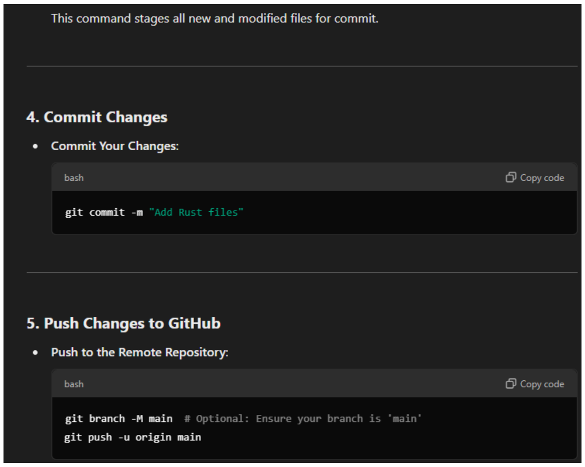
- **Option 2: Clone an Existing Repository:**

```bash
git clone https://github.com/username/repository.git
cd repository
```

3. Add Rust Files to Your Repository

- **Write Your Rust Code:** Ensure your Rust files are in the repository directory. You can use tools like Visual Studio Code or any text editor to write your Rust files.
- **Track the Files with Git:**

```bash
git add .
```



This command stages all new and modified files for commit.

4. Commit Changes

- **Commit Your Changes:**

```bash
git commit -m "Add Rust files"
```

5. Push Changes to GitHub

- **Push to the Remote Repository:**

```bash
git branch -M main  # Optional: Ensure your branch is 'main'
git push -u origin main
```

The dataset I used for this project is found online as House Price Prediction Dataset on Kaggle. After downloading and moving it to my project's directory, I used a similar code structure as in the lecture to first read the data and store it in the DirtyHouseRecord struct. I then parsed and converted it to a clean version and stored it in the CleanHouseRecord. This code was put into the model.rs module.

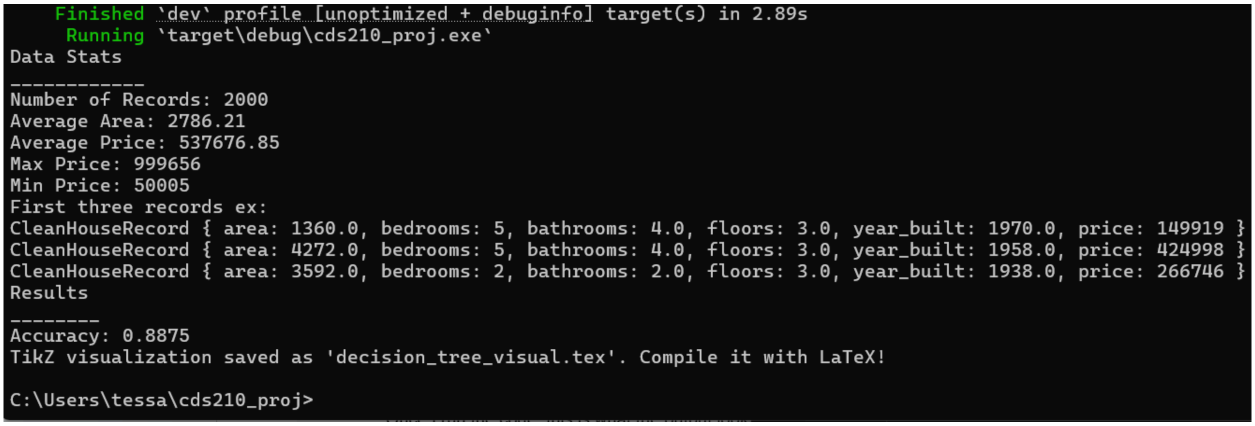
I then did some exploratory analysis on the dataset in the data.rs module such as finding the number of records to make sure all records were being processed properly, and calculating the averages of areas and prices. The functions process_csv and explore_data also calculate and print the maximum and minimum prices.

Then, in main.rs, I import the relevant crates and modules, such as linfa_trees, ndarray, and the data and model modules. The main function has the path to the file location and cleans it using process_csv. Then, the explore_data function prints the stats from the data.rs file as well as the first three records to make sure the correct numbers are being used. After that, I created two vectors to hold feature values and labels from the data for the decision tree, and then extracted these values and labeled them based on if the price is greater or less than 500,000. Then, I converted both vectors into 2d arrays and created a dataset with the corresponding features and labels.

After that, I trained the decision tree and set the max depth to 20 in the final version. Earlier, when I had tried lower values like 5, the accuracy of the decision tree had been around 60%. The highest I got it to was around 94.5% with a maximum depth of 35ish, but that would have made the decision tree too complex. Then, I used decsion_tree.predict to make predictions for the data, and then pred.confusion_matrix(&dataset) to make a confusion matrix to use to calculate the accuracy.

The last portion of code creates a TikZ file that can be used in LaTeX to create a visualization as shown in an example in the lecture notes. As for tests, the test_process_csv function makes sure that the process_csv doesn't leave anything blank, and the test_clean_csv function tests the clean_csv function to make sure that the output is correct using a sample of dirty data.

Once I run the code, this is what the output looks like in my terminal:



```
    Finished `dev` profile [unoptimized + debuginfo] target(s) in 2.89s
     Running `target\debug\cds210_proj.exe`
Data Stats
------------
Number of Records: 2000
Average Area: 2786.21
Average Price: 537676.85
Max Price: 999656
Min Price: 50005
First three records ex:
CleanHouseRecord { area: 1360.0, bedrooms: 5, bathrooms: 4.0, floors: 3.0, year_built: 1970.0, price: 149919 }
CleanHouseRecord { area: 4272.0, bedrooms: 5, bathrooms: 4.0, floors: 3.0, year_built: 1958.0, price: 424998 }
CleanHouseRecord { area: 3592.0, bedrooms: 2, bathrooms: 2.0, floors: 3.0, year_built: 1938.0, price: 266746 }
Results
--------
Accuracy: 0.8875
TikZ visualization saved as 'decision_tree_visual.tex'. Compile it with LaTeX!

C:\Users\tessa\cds210_proj>
```

The TikZ file is also automatically saved to the project folder.