UNIKEY: lshe0103

1. Topic and research question

In this research project, the central research question is: “How does the forest cover of a patch depend on the relationship between soil type and wilderness areas?”. This research aims to explore the relationship between forest cover, soil type, and wilderness areas, with a specific focus on determining whether certain tree species are better suited to specific soil types or if they are just more likely to thrive in particular wilderness areas (Since different wildness area have different elevation, humidness, etc.). Understanding this relationship is crucial. By analyzing how these factors influence forest cover, we can better predict which tree species will thrive in specific regions. This is especially important in the context of ecological restoration and conservation efforts. This research is particularly valuable for environmental protection agencies and botanists. For environmental protection agencies, the research can provide clear guidelines for planting trees in specific wilderness areas, maximizing tree survival rates by choosing species that are best suited to the local soil conditions. For botanists, this study offers a comprehensive understanding of tree species distribution across different wilderness areas, enhancing knowledge of plant ecology. Additionally, by analyzing soil types, botanists can predict which tree species might be discovered in previously unexplored wilderness areas, potentially leading to new findings in biodiversity.

2. Data description

The dataset contains data collected from four wilderness areas: Neota, Rawah, Comanche Peak, and Cache la Poudre. Within each wilderness area, multiple patches are examined, with each patch having a uniform size. The dataset consists of 30,860 instances (i.e., patches) and 56 attributes, including an attribute representing the index of each instance. These attributes describe various environmental characteristics of the patches, such as elevation, slope, aspect, soil type, and wilderness area classification, which may be relevant to determining the forest cover type. A comprehensive data dictionary is provided in Appendix 1, detailing each attribute’s description, data type (e.g., integer, float, string), and valid range. Additionally, some columns may contain invalid or inconsistent values, which will be addressed in the data cleaning section.

3.1 Data ingestion

To ingest this CSV dataset, I used Python’s Pandas library and stored it as a Pandas DataFrame. Specifically, I employed the pandas.read\_csv() function to load the dataset from "forest\_cover.csv”, meaning all data, including the index column, is stored accordingly. Since the dataset is contained within a single CSV file, only one DataFrame is created for this analysis. Pandas automatically detects the data type of each column based on its values. For example, columns containing only integer values are assigned to the int type, while columns containing both integers and floats are assigned to the float type. Certain columns have different data types due to missing values (NaN): for instance, "Soil\_Type7" to "Soil\_Type11" contain only integers, so their column type remains int, whereas other "Soil\_Type" columns contain NaN values, causing Pandas to interpret them as float. The dataset consists of 56 attributes (including index column), which can be broadly categorized into interval, nominal, and ratio attributes. The interval attribute "Aspect" is stored as a float, while ratio attributes such as "Elevation", "Slope", "Horizontal\_Distance\_To\_Hydrology", "Vertical\_Distance\_To\_Hydrology", "Horizontal\_Distance\_To\_Roadways", "Hillshade\_9am", "Hillshade\_Noon", "Hillshade\_3pm", and "Horizontal\_Distance\_To\_Fire\_Points" are stored as either integers or floats, depending on the specific data. Nominal attributes include "Soil\_Type1" through "Soil\_Type40" and the wilderness area columns "Neota", "Rawah", "Comanche Peak", and "Cache la Poudre" which follow a one-hot encoding scheme—only one column in each group can have a value of 1, with all others remaining 0. Another nominal attribute, "Forest\_Cover" is stored as an object type because Pandas recognizes all non-numeric columns as object.

3.2 Data quality assurance and cleaning

The dataset has several data quality issues that require thorough cleaning. Except for the response variable "Forest Cover", all other columns contain missing values. Additionally, the "Soil\_Type" columns use one-hot encoding, which would be better represented in a single column rather than multiple columns. The same approach applies to the "Neota", "Rawah", "Comanche Peak", and "Cache la Poudre" columns. Several columns, such as "Hillshade\_9am", "Hillshade\_Noon", "Hillshade\_3pm", "Horizontal\_Distance\_To\_Hydrology", "Horizontal\_Distance\_To\_Roadways", "Horizontal\_Distance\_To\_Fire\_Points", and "Aspect" contain values that exceed valid ranges, such as negative horizontal distances, which do not make sense.

To address the missing values in the "Elevation" column, I first analyzed the average elevation for different wilderness areas using a combination of dropna(), groupby(), and mean() functions, counting only rows with valid elevation data. I then filled the missing values by using the average elevation for the corresponding wilderness area. This approach makes sense because wilderness areas generally have minimal elevation differences, thus preserving the original trends in the data. For the missing values in the "Soil\_Type" columns, I first observed that only one column per row could have a value of 1, ensuring that each row belongs to a single soil type. In such cases, I filled the NaN with 0 when a row already had a soil type. For rows lacking a soil type, I categorized the data based on wilderness areas, as each wilderness area has distinct soil types. I calculated the distribution of soil types in each wilderness area and filled the missing values according to the probabilities of these distributions. This ensured that the data pattern remained intact, and I used the random.choices() function to assign soil types. This method is valid as most wilderness areas have a majority of valid rows. Once the missing values were filled, I combined the individual soil type columns into one "Soil\_Type" column with an int datatype, which ranges from 1 to 40, for ease of analysis. For missing values in the "Hillshade\_9am”, "Hillshade\_Noon”, and "Hillshade\_3pm" columns, I filled the missing values by using the mean of the other two columns if only one value was missing. If two values were missing, I used the remaining value to fill the others. For rows with all three missing values, I dropped them, as they accounted for only 320 rows. This approach is reasonable because the hill shade values do not change much from 9 am to 3 pm. I also replaced illegal values in these columns, such as negative numbers and values greater than 255, by setting them to 0 and 255, respectively. Regarding missing values in the "Horizontal/Vertical\_Distance\_To\_XXXX” columns, I dropped the "Vertical\_Distance\_To\_Hydrology" column after analyzing its distribution, as it showed minimal variation. The remaining columns were filled with their respective means directly, as these columns did not seem strongly related to wilderness areas. I used the iterrows() function and loc[] method to fill missing values and filtered out negative values by applying the abs() function. This ensured that all horizontal distances were positive, preserving the logical consistency of the data. For the "Aspect" and "Slope" columns, I decided to drop them as the mean values for each tree type were very similar, and their removal did not significantly affect the dataset. I used the drop() function to eliminate these columns. For the "Neota”, "Rawah”, "Comanche Peak”, and "Cache la Poudre" columns, which were one-hot encoded, I simply filled missing values with 0, as each row had exactly one of these columns with a value of 1. I also combined these columns into one column named "Wilderness\_Area”, with a string datatype. The values in this column can only be one of "Neota”, "Rawah”, "Comanche Peak”, or "Cache la Poudre." In summary, I employed a combination of analysis techniques, including grouping, probability distributions, and random selection, to address missing values and correct invalid entries. I also used various functions like dropna(), groupby(), apply(), and iterrows() to ensure the data was cleaned while maintaining its integrity for further analysis.

4. Exploratory Data Analysis

Plot 1 Plot 2

A graph of different soil types

AI-generated content may be incorrect. A graph of different colors

AI-generated content may be incorrect.

To investigate the research question, which explores the relationship between forest cover (tree species), soil types, and wilderness areas, I used stacked bar plots to reveal potential trends between these variables. I chose this method because all three variables are nominal, and it is essential to consider the count of each nominal type. The y-axis represents the count, and the stacked bar plots visually convey the distribution of tree species across different soil types and wilderness areas.

By "related," I mean that a tree species occupies approximately 50% or more of the total in a specific soil type, indicating a strong association.

From Plot 1, Lodgepole Pine is most associated with soil types 9, 10, 11, 12, 13, 29, and 30. In Plot 2, Lodgepole Pine primarily separates in the Comanche Peak and Rawah wilderness areas, which are dominated by soil types 12, 23, 29, 30, 31, 32, and 33. Common soil types between the two plots include 12, 29, and 30. Interestingly, Lodgepole Pine is less associated with soil types in Comanche Peak, where the top four soil types (32, 33, 31, and 23) do not overlap with those in Plot 1. One possible explanation is that each wilderness area only lists the top four soil types, and less common soil types like 13, ranked fifth with 1038 instances (just below the fourth-ranked 1155), are not represented despite their strong association with Lodgepole Pine. This absence could account for the lack of Lodgepole Pine-related soil types in Comanche Peak (Plot 2).

Next, for Spruce/Fir, Plot 1 shows associations with soil types 19, 20, 22, 23, 24, 29, 30, 31, 32, 33, and 35. In Plot 2, Spruce/Fir is mostly distributed in Comanche Peak (with top four soil types 32, 33, 31, and 23), Rawah wilderness (dominated by soil types 29, 30, 23, and 12), and Neota (with top four soil types 23, 22, 32, and 33). The overlap between these regions includes soil types 22, 23, 29, 30, 31, 32, and 33, which align with most of the soil types associated with Spruce/Fir.

For Ponderosa Pine, in Plot 1, it is associated with soil types 1, 2, 3, 4, 5, and 6. In Plot 2, it is mainly related to Cache la Poudre, where the top four soil types are 10, 6, 1, and 3. The common soil types between the two plots are 6, 1, and 3, which align with most of the soil types associated with Ponderosa Pine.

In summary, there are variations across different wilderness areas, the data suggests that soil type plays a more significant role than other environmental factors in determining tree species.. However, the dataset has some significant issues, with Lodgepole Pine and Spruce/Fir dominating the distribution. This imbalance could lead to bias when building a machine learning model. To address this issue, one potential approach is to assign more weight to underrepresented tree species during model training. Additionally, techniques such as resampling (either over-sampling the minority class or under-sampling the majority class) or generating synthetic samples using methods like SMOTE (Synthetic Minority Over-sampling Technique) can help balance the dataset. Another option is to use ensemble methods like Random Forests or XGBoost, which are designed to handle imbalanced data. These strategies will be explored further in Assignment 2.

Forest\_Cover

UNIKEY: lshe0103

1. Topic and research question

Forests play a crucial role in maintaining ecological balance by providing habitats for wildlife, regulating climate, and preventing soil erosion. Understanding the factors that influence forest cover is essential for managing ecosystems and supporting biodiversity. In this research project, the research question is: **“How does forest cover depend on the relationship between soil type and elevation?”** This study investigates whether certain tree species are primarily suited to specific soil types or if elevation also plays a significant role in determining forest cover.  
 Understanding this relationship is critical for predicting which tree species will thrive in particular regions, thus contributing to more effective ecological restoration and conservation efforts. This research holds significant value for both environmental protection agencies and botanists. For environmental agencies, the findings can inform tree planting strategies, helping to maximize survival rates by selecting species that are best adapted to local soil conditions. For botanists, this study deepens our understanding of tree species distribution across different wilderness areas, enhancing our knowledge of plant ecology. Additionally, by analyzing soil types, botanists may be able to predict which tree species might be found in previously unexplored regions, contributing to the discovery of new biodiversity.

2. Data description

The dataset contains data collected from four wilderness areas: Neota, Rawah, Comanche Peak, and Cache la Poudre. Within each wilderness area, multiple patches are examined, with each patch having a uniform size. The dataset consists of 30,860 instances (i.e. patches) and 56 attributes, including an attribute representing the index of each instance. These attributes describe various environmental characteristics of the patches, such as elevation, slope, aspect, soil type, and wilderness area classification, which may be relevant to determining the forest cover type. A comprehensive data dictionary is provided in Appendix 3, detailing each attribute’s description, data type (e.g., integer, float, string), and valid range. Additionally, some columns may contain invalid or inconsistent values, which will be addressed in the data cleaning section.Lastly, for convenience, I have also included two new columns from the cleaned dataset in the data dictionary and highlighted them in bold.

3.1 Data ingestion

To ingest this CSV dataset, I used Python’s Pandas library and stored it as a Pandas DataFrame. Specifically, I employed the pandas.read\_csv() function to load the dataset from "forest\_cover.csv", ensuring that all data, including the index column, was properly stored. Since the dataset is contained within a single CSV file, only one DataFrame was created for this analysis. Pandas automatically detects the data type of each column based on its values. For instance, columns containing only integer values are assigned the int type, while those containing both integers and floats are assigned the float type. Certain columns exhibit different data types due to missing values (NaN); for example, "Soil\_Type7" to "Soil\_Type11" contain only integers and remain as int, whereas other "Soil\_Type" columns contain NaN values, causing Pandas to interpret them as float. The dataset consists of 56 attributes (including the index column), which can be broadly categorized into interval, nominal, and ratio attributes. The interval attribute "Aspect" is stored as a float, while ratio attributes such as "Elevation," "Slope," "Horizontal\_Distance\_To\_Hydrology," "Vertical\_Distance\_To\_Hydrology," "Horizontal\_Distance\_To\_Roadways," "Hillshade\_9am," "Hillshade\_Noon," "Hillshade\_3pm," and "Horizontal\_Distance\_To\_Fire\_Points" are stored as either int or float, depending on the specific data. Nominal attributes include "Soil\_Type1" through "Soil\_Type40" and the wilderness area columns ("Neota," "Rawah," "Comanche Peak," and "Cache la Poudre"). These attributes follow a one-hot encoding scheme, where only one column in each group can have a value of 1 while all others remain 0. Due to the presence of NaN values, Pandas may store them as either int or float. Another nominal attribute, "Forest\_Cover," is stored as an object type, as Pandas recognizes non-numeric columns as object by default.

3.2 Data quality assurance and cleaning

The dataset contains several data quality issues that need to be addressed. With the exception of the response variable "Forest Cover", the wilderness area columns ("Neota," "Rawah," "Comanche Peak," and "Cache la Poudre") and the columns “Soil\_Type7” to “Soil\_Type11”, all other columns contain missing values. Additionally, the "Soil\_Type" columns use one-hot encoding, which is inefficient and could be represented more effectively in a single column instead of 40 separate columns. The same issue applies to the "Neota", "Rawah", "Comanche Peak", and "Cache la Poudre" columns. Several columns, such as "Hillshade\_9am", "Hillshade\_Noon", "Hillshade\_3pm", "Horizontal\_Distance\_To\_Hydrology", "Horizontal\_Distance\_To\_Roadways", "Horizontal\_Distance\_To\_Fire\_Points", and "Aspect", contain values that fall outside of valid ranges. For example, negative horizontal distances are present, which are not meaningful or valid in this context. Finally, approximately 80% of the data is focused on the “Spruce/Fir” and “Lodgepole Pine” forest cover types. This imbalance may introduce bias into the analysis, potentially limiting the generalizability of findings for other tree species.  
 To address the missing values in the "Elevation" column, I first analyzed the average elevation for different wilderness areas using a combination of dropna(), groupby(), and mean() functions, counting only rows with valid elevation data. I then filled the missing values by using the average elevation for the corresponding wilderness area. This approach makes sense because wilderness areas generally have minimal elevation differences, thus preserving the original trends in the data.   
 For the missing values in the "Soil\_Type" columns, I first observed that only one column per row could have a value of 1, ensuring that each row belongs to a single soil type. In such cases, I filled the NaN with 0 when a row already had a soil type. For rows lacking a soil type, I categorized the data based on wilderness areas, as each wilderness area has a distinct soil type.I calculated the distribution of soil types in each wilderness area and then filled in the missing values based on the wilderness area type of the current row. This ensured that the data pattern remained intact, and I used the random.choices() function to assign soil types. This method is valid as most wilderness areas have a majority of valid rows. Once the missing values were filled, I combined the individual soil type columns into one "Soil\_Type" column with an int datatype, which ranges from 1 to 40, for ease of analysis.   
 For missing values in the "Hillshade\_9am”, "Hillshade\_Noon”, and "Hillshade\_3pm" columns, I filled in the missing values by using the mean of the other two columns if only one value was missing. If two values were missing, I used the remaining value to fill the others. For rows with all three missing values, I dropped those row, as they accounted for only 320 rows. This approach is reasonable because the hill shade values do not change much from 9 am to 3 pm. I also replaced illegal values in these columns, such as negative numbers and values greater than 255, by setting them to 0 and 255, respectively.   
 I dropped the "Vertical\_Distance\_To\_Hydrology" column after analyzing its distribution, as it showed minimal variation. Regarding missing values in the "Horizontal/Vertical\_Distance\_To\_XXXX” columns, they were filled with their respective means directly, as these columns did not seem strongly related to wilderness areas. I used the iterrows() function and loc[] method to fill missing values and change negative values to positive by applying the abs() function. This ensured that all horizontal distances were positive, preserving the logical consistency of the data.   
 For the "Aspect" and "Slope" columns, I decided to drop them as the mean values for each tree type were very similar, and their removal did not significantly affect the dataset. I used the drop() function to eliminate these columns.   
 For the "Neota”, "Rawah”, "Comanche Peak”, and "Cache la Poudre" columns, which were one-hot encoded, I simply filled missing values with 0, as each row had exactly one of these columns with a value of 1. I also combined these columns into one column named "Wilderness\_Area”, with a string datatype. The values in this column can only be one of "Neota”, "Rawah”, "Comanche Peak”, or "Cache la Poudre" .  
 In summary, I utilized a combination of analytical techniques, including grouping, probability distributions, and random selection, to handle missing values and correct invalid entries. Additionally, I applied various functions such as dropna(), groupby(), apply(), and iterrows() to clean the data while preserving its integrity for further analysis. Lastly, I retained columns that were not directly relevant to my research question to maintain as much data as possible for Assignment 2.

4. Exploratory Data Analysis

A graph of different types of soil type

AI-generated content may be incorrect.

    To investigate the research question, I first explore the relationship between forest cover (tree species) and soil type. In Plot 1, I used stacked bar plots to illustrate the distribution of forest covers across different soil types. This method was chosen because both variables are nominal, and it is important to visualize the count of each type. The y-axis represents the count, and the stacked bars effectively show how tree species are distributed across soil types and wilderness areas. From this plot, we can conclude that tree species do show a strong relationship with soil types. For example, Lodgepole Pine tends to prefer soil types 10, 11, 12, 13, 29, and 30, while Spruce/Fir is more commonly found in soil types 19, 20, 22, 23, 24, 29, 31, 32, and 33. Ponderosa Pine prefers soil types 1, 2, 4, and 6, and Krummholz is typically found in soil types 38, 39, and 40. These patterns indicate that certain tree species are indeed more suited to specific soil types. However, for other tree species with limited data, we cannot draw clear conclusions from the plot. It's also important to note that the majority of the data is focused on Lodgepole Pine and Spruce/Fir, so these conclusions are particularly reliable for these two species, while the results for other species may not be as accurate.

A diagram of a diagram

AI-generated content may be incorrect.

Now, let’s use Plot 2 to explore the relationship between elevation and different tree species. In this plot, I used a box plot to represent the relationship between a ratio variable (elevation) and a nominal variable (forest cover type). The X-axis represents the forest cover type (tree species) with its Coefficient of Variation (CV), and the Y-axis represents elevation. From the plot, we can observe that most tree species are not concentrated around a specific elevation, as indicated by the wide distance between Q1 and Q3 for each species. The CV further supports this, with all CV values greater than 20%, indicating that the elevation data is spread out and not concentrated around the mean. However, we do see that most tree species tend to be clustered between elevations of 2200 and 3600 meters, suggesting that while there isn't a strong correlation, elevation may still play a role in tree distribution. In conclusion, while there is some relationship between tree species and elevation, it is not strong enough to make definitive claims, and elevation alone does not strongly determine forest cover type.

In summary, there is a strong relationship between soil type and forest cover type, as demonstrated in Plot 1, where specific tree species are clearly associated with certain soil types. However, the relationship between forest cover and elevation in Plot 2 is less apparent. Trees can thrive across a wide range of elevations, and some outliers even exceed this range, indicating that elevation does not have a strong influence on forest cover type.

It’s important to note that due to the imbalance in the dataset, with Lodgepole Pine and Spruce/Fir dominating the distribution, the conclusions drawn from this analysis may be biased. Since the majority of the data is focused on these two tree species, the findings regarding other tree types may not be as reliable. This bias could affect subsequent analysis or model-building tasks. For instance, the distribution of soil types and elevations may not be representative for less frequent tree species, potentially leading to overgeneralized conclusions. To address this imbalance in the machine learning model for Assignment 2, one potential solution is to assign higher weights to the underrepresented tree species during model training. However, even with this approach, certain tree species that prefer specific soil types, which are not well represented in the dataset, may still not be adequately covered. While weighting can help mitigate some of the imbalance, it may not fully resolve the issue if the data for particular soil types or tree species remains sparse.

Appendix 1

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Name | Description | Data Type | Range |
| Elevation | Elevation of current patch in meters. | float64 | any float number |
| Aspect | Describes the orientation of a slope relative to cardinal directions. | float64 | [0:360] |
| Slope | The steepness or incline of the patch, measured in degrees. | float64 | any float number |
| Horizontal\_Distance\_To\_Hydrology | Horizontal distance to nearest surface water features, measured in meters. | float64 | non-negative float number |
| Vertical\_Distance\_To\_Hydrology | Vertical distance to nearest surface water features, measured in meters. | float64 | any float number |
| Horizontal\_Distance\_To\_Roadways | Horizontal distance to the nearest roadway, measured in meters. | float64 | non-negative float number |
| Horizontal\_Distance\_To\_Fire\_Points | Horizontal distance to nearest wildfire ignition points, measured in meters. | float64 | non-negative float number |
| Hillshade\_9am | Hill shade index at 9am, summer solstice. 0 means complete darkness, 255 means full sunlight. | float64 | [0:255] |
| Hillshade\_Noon | Hill shade index at 9am, summer solstice. 0 means complete darkness, 255 means full sunlight. | float64 | [0:255] |
| Hillshade\_3pm | Hill shade index at 9am, summer solstice. 0 means complete darkness, 255 means full sunlight. | float64 | [0:255] |
| Soil\_Type1 & … & Soil\_Type40 | These 40 columns represent the soil type of the current patch. Each patch can only have one type of soil, meaning only one of the Soil\_Type columns can be set to 1 at a time, with all other columns set to 0. | float64 | either 0 or 1 |
| Forest\_Cover | The response variable, representing the type of forest cover in the patch. | object | It should be one of the following: 'Spruce/Fir', 'Lodgepole Pine', 'Ponderosa Pine', 'Aspen', 'Douglas-fir', 'Krummholz', or 'Cottonwood/Willow'. |
| Neota | Represents the wilderness area to which the current patch belongs. Only one of the wilderness area columns can be set to 1 at a time, indicating the specific wilderness area for each patch. All other columns will be set to 0. | int64 | either 0 or 1. |
| Rawah | Represents the wilderness area to which the current patch belongs. Only one of the wilderness area columns can be set to 1 at a time, indicating the specific wilderness area for each patch. All other columns will be set to 0. | int64 | either 0 or 1. |
| Comanche Peak | Represents the wilderness area to which the current patch belongs. Only one of the wilderness area columns can be set to 1 at a time, indicating the specific wilderness area for each patch. All other columns will be set to 0. | int64 | either 0 or 1. |
| Cache la Poudre | Represents the wilderness area to which the current patch belongs. Only one of the wilderness area columns can be set to 1 at a time, indicating the specific wilderness area for each patch. All other columns will be set to 0. | int64 | either 0 or 1. |