Title: Project "Street Light Classification and Anomaly Detection: Enhancing Infrastructure Planning and Maintenance with Machine Learning"

Name: Gurminder Singh Badwal

Student Id: 041116396

Date of Submission: 29, March 2024

Contents

[Introduction: 5](#_Toc162628079)

[Business Understanding: 6](#_Toc162628080)

[1. Determine Business Objectives: outlier and classification. 6](#_Toc162628081)

[2. Assessing the Situation: 6](#_Toc162628082)

[3. Data Mining Goals: 6](#_Toc162628083)

[4. Produce Project Plan: 7](#_Toc162628084)

[Data Understanding: 7](#_Toc162628085)

[1. Collect Initial Data: 7](#_Toc162628086)

[2. Describe Data: 7](#_Toc162628087)

[**1.** **Pole ID:** 7](#_Toc162628088)

[Unique identifier for each light pole. 7](#_Toc162628089)

[It provides specific identification information. 7](#_Toc162628090)

[**2.** **Pole Installation Date:** 7](#_Toc162628091)

[It provides unique information about the installation timing. 8](#_Toc162628092)

[**3.** **Light Installation Date:** 8](#_Toc162628093)

[Specifies the date when the streetlight was installed. 8](#_Toc162628094)

[It provides unique information about the installation timing and is not redundant. 8](#_Toc162628095)

[**4.** **Light Head Style:** 8](#_Toc162628096)

[Describes the style or design of the head or housing of the street light fixture. 8](#_Toc162628097)

[It provides unique information about the physical appearance of the light fixture and is not redundant. 8](#_Toc162628098)

[**5.** **Refractor:** 8](#_Toc162628099)

[Represent the refractor type or component used to refractor light according to the distribution patterns. 8](#_Toc162628100)

[**6.** **Light Type:** 8](#_Toc162628101)

[Indicates the type or category of the streetlight (e.g., streetlight, pedestrian light, lawn light, wall-mounted light, solar street light, recreational light). 8](#_Toc162628102)

[**7.** **Bracket Length:** 8](#_Toc162628103)

[Represents the length of the bracket used to support or mount the street light fixture. 8](#_Toc162628104)

[**8.** **Pole Height:** 8](#_Toc162628105)

[Specifies the height of the light pole supporting the street light fixture. 8](#_Toc162628106)

[**9.** **Pole Type:** 8](#_Toc162628107)

[Indicates the material or type of the light pole. 9](#_Toc162628108)

[**10.** **Number of Lights:** 9](#_Toc162628109)

[Indicates the number of lights associated with the pole. 9](#_Toc162628110)

[**11.** **Light Source:** 9](#_Toc162628111)

[Indicates the source of light used in the street light fixture. 9](#_Toc162628112)

[**12.** **Wattage:** 9](#_Toc162628113)

[Represents the electrical power consumption of the street light fixture, measured in watts. 9](#_Toc162628114)

[**13.** **Year 1 Dimming Wattage:** 9](#_Toc162628115)

[Indicates the wattage of the street light fixture in the first year of operation when dimmed. 9](#_Toc162628116)

[**14.** **Street Name:** 9](#_Toc162628117)

[Specifies the name of the street where the streetlight is located. 9](#_Toc162628118)

[**15.** **Energy Jurisdiction:** 9](#_Toc162628119)

[Indicates the jurisdiction or authority responsible for energy management or regulation. 9](#_Toc162628120)

[**16.** **CREATED\_DATE:** 9](#_Toc162628121)

[Specifies the date when the record was created. 9](#_Toc162628122)

[**17.** **MODIFIED\_DATE:** 9](#_Toc162628123)

[Indicates the date when the record was last modified. 9](#_Toc162628124)

[3. Explore Data: 10](#_Toc162628125)

[4. Verify Data: 11](#_Toc162628126)

[Data Preparation: 13](#_Toc162628127)

[1. Select Data: 13](#_Toc162628128)

[2. Clean Data: 16](#_Toc162628129)

[3. Construct Data: 17](#_Toc162628130)

[4. Integrate Data: 21](#_Toc162628131)

[5. Format Data: 21](#_Toc162628132)

[Modelling: 23](#_Toc162628133)

[Outlier Detection 23](#_Toc162628134)

[ISF 23](#_Toc162628135)

[LOF: 24](#_Toc162628136)

[Classification: 26](#_Toc162628137)

[KNN: K- Nearest Neighbors simple and intuitive algorithm used for both classification and regression tasks, the prediction for a new data point is determined by the majority class (for classification) or the average value (for regression) of its k nearest neighbors in the feature space, KNN's performance can be sensitive to the choice of the distance metric and the value of k. 26](#_Toc162628138)

[Decision Tree: 31](#_Toc162628139)

[Random Forest: 41](#_Toc162628140)

[Conclusion: 48](#_Toc162628141)

[References: 50](#_Toc162628142)

Top of FormProject: "Street Light Classification and Anomaly Detection: Enhancing Infrastructure Planning and Maintenance with Machine Learning"

Introduction: In this project, our objective is to answer, "What are the key contributing factors influencing the classification of street lights into different types based on various attributes, and how can outlier detection techniques help identify anomalies within the dataset?” By addressing these objectives, we aim to provide municipalities and transportation departments with valuable tools for better infrastructure planning, maintenance prioritization, and resource allocation.

The dataset we are working with contains comprehensive information about street lights, including geographic coordinates, physical attributes (such as head style, bracket length, and pole height), operational characteristics (such as wattage and light source), and additional contextual data. Through rigorous analysis of these attributes, we seek to construct classification models that accurately predict the type of street light fixture based on its features. Additionally, we aim to employ outlier detection techniques to identify anomalies or irregularities in the dataset that may warrant further investigation.

This project holds significant potential to yield actionable insights for urban planning and infrastructure management. By developing accurate classification models and detecting outliers effectively, municipalities can optimize maintenance schedules, prioritize infrastructure upgrades, and enhance overall efficiency in managing street lighting networks. Furthermore, the findings from this analysis can inform decision-making processes related to energy conservation, public safety, and urban development.

In the subsequent sections, we will delve into our approach, encompassing data understanding, data preparation, model development, and evaluation. Through this endeavor, we aim to demonstrate the value of data-driven methodologies in enhancing urban infrastructure management and fostering sustainable, well-illuminated communities.

# Business Understanding:

## 1. Determine Business Objectives: outlier and classification.

This Document details about prediction of type of streetlight (SL: Street Light, Pedestrian Light, Lawn Light, Wall-Mounted Light, Solar Street Light, Recreational Light) based on various attributes such as head style, bracket length, pole height, pole type, wattage, and aims to detect outliers within the street light dataset, which could indicate potential anomalies or errors in data recording, as well as identify street lights that may require immediate attention due to unusual characteristics or maintenance needs. Predicting the type of street light, the street lighting designers, manufacturers, and purchase decision makers can make better choices related to profit and growth.

## 2. Assessing the Situation:

The Street Light dataset which we have consists of various attributes related to XY point geometry, style or design of the head of street light fixture, light type, bracket length, pole height, pole type wattage, light source etc. As per the situation, the data is not fully complete, there are missing instances from the attributes, so the selection and exclusion shall also happen according to this in the steps of Data Preparation. We have access to efficiently use the machine learning algorithms such as decision tree, KNN and outlier detection methods like LOF and ISF through software known as RapidMiner.

## 3. Data Mining Goals:

* The goal is to provide accurate predictions of street light types, which can assist street lighting designers, manufacturers, and decision-makers in making informed choices related to infrastructure planning, maintenance, and procurement.
* And to enhance data quality and identify potential maintenance issues or anomalies in street lighting infrastructure, enabling proactive maintenance and resource allocation decisions.
* Utilize machine learning algorithms such as decision trees and KNN for classification, LOF and ISF for outlier detection.

## 4. Produce Project Plan:

The project plan:

1. The Data should be ready for testing out the algorithms and detections beforehand in the phase of Data Preparation so that there is no issue while going through the phases of modelling.
2. All the classification and outlier detection methods would be implemented by me as an author of this report.
3. Evaluation will be happened based on both the KNN and Decision tree classification methods and check the anomalies presented by the outlier detections which are affecting the outcomes or needed to know to make confirmed decisions about finding the type of lights, which is what our overall aim is and that is what this project is all about.
4. After evaluation the final project with all the outcomes will be shown via presentations.

# Data Understanding:

## Collect Initial Data:

Initial data is collected form the open Ottawa platform where we download and analyze the data according to our project needs. And data for this is about Street Lights with more than 70000 instances.

## Describe Data:

In total there are 74,031 instances of the dataset but for this report 10000 instances are selected for performing all the operations mentioned before. The information about attributes is as below with their datatypes:

## **Pole ID:**

Type: Character string

## Unique identifier for each light pole.

## It provides specific identification information.

## **Pole Installation Date:**

Type: Date

Indicates the date when the light pole was installed.

## It provides unique information about the installation timing.

## **Light Installation Date:**

Type: Date

## Specifies the date when the streetlight was installed.

## It provides unique information about the installation timing and is not redundant.

## **Light Head Style:**

Type: Character string

## Describes the style or design of the head or housing of the street light fixture.

## It provides unique information about the physical appearance of the light fixture and is not redundant.

## **Refractor:**

Type: Character string

## Represent the refractor type or component used to refractor light according to the distribution patterns.

## **Light Type:**

Type: Character string

## Indicates the type or category of the streetlight (e.g., streetlight, pedestrian light, lawn light, wall-mounted light, solar street light, recreational light).

## **Bracket Length:**

Type: Real (Double-precision decimal number)

## Represents the length of the bracket used to support or mount the street light fixture.

## **Pole Height:**

Type: Real (Double-precision decimal number)

## Specifies the height of the light pole supporting the street light fixture.

## **Pole Type:**

Type: Character string

## Indicates the material or type of the light pole.

## **Number of Lights:**

Type: Integer

## Indicates the number of lights associated with the pole.

## **Light Source:**

Type: Character string

## Indicates the source of light used in the street light fixture.

## **Wattage:**

Type: Character string

## Represents the electrical power consumption of the street light fixture, measured in watts.

## **Year 1 Dimming Wattage:**

Type: Character string

## Indicates the wattage of the street light fixture in the first year of operation when dimmed.

## **Street Name:**

Type: Character string

## Specifies the name of the street where the streetlight is located.

## **Energy Jurisdiction:**

Type: Character string

## Indicates the jurisdiction or authority responsible for energy management or regulation.

## **CREATED\_DATE:**

Type: Date

## Specifies the date when the record was created.

## **MODIFIED\_DATE:**

Type: Date

## Indicates the date when the record was last modified.

## Explore Data:

This chart is the class labeled chart of Light type attribute which shows the distribution of each type of light among the instances provided.

A graph with a bar and text

Description automatically generated

After analyzing the data, the understanding is confirmed that type of light attribute, the SL type which is streetlight is more using light type in all over the Ottawa streets. Whereas other types are not that used according to the specific circumstances of streets such as waterproof lights are used there where there are fountains, or high chance of water revelation on surface.

Other than this Light Type and Other Attributes, for example, the other attributes (e.g., head style, bracket length, pole height, pole type, wattage) may vary depending on the type of light, so certain types of streetlights might require taller poles or specific head styles according to the light.

Wattage and Light Source, there may be a relationship between the wattage of the streetlight and its light source. Different light sources (e.g., LED, High Pressure Sodium) may have different wattage requirements to achieve a certain level of illumination.

Pole Height and Pole Type: Pole height and pole type could be related attributes. For instance, certain pole types might be designed for specific heights or weight-bearing capacities.

And also, I explored of Street Names attribute that it is imperative attribute for consideration as part for modelling because it shows the regions where the most amount of light are installed and what are the types of those lights as we can see if all those lights are of LED type or not as the aim for Ottawa lights is to converts them to LED type. And these insights are important while classifying as these independent variables are what going to show their presence on dependent variable by their coefficient to classify each type.

So, these are what we have explored through this exploration about the possible relationships between the attributes.

## Verify Data:

Data is not complete as previously mentioned there are attributes with missing values such as light type and refractor attribute, there is some noise present in attribute head style where some of the values are not common ones, which needs to focus on but needed for selection purposes.

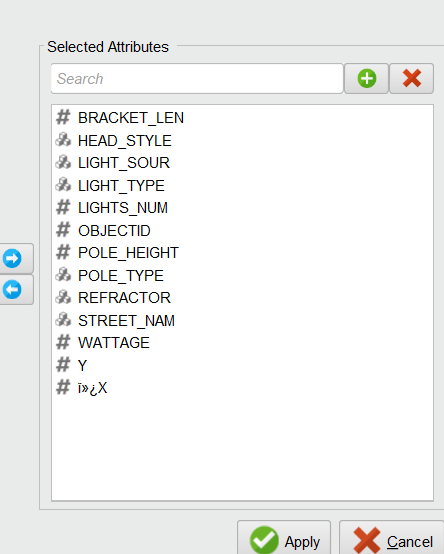
Data has no duplicate values but yes data is not complete, and I am only considering the specific parts of data which I mentioned in the phase of selecting the data.A screenshot of a computer

Description automatically generated

Additional updates are also pending which contributors are solving on quarterly basis, and one point to be noted is that not all the Ottawa street lighting is shifted on LED (Light Emitting Diode) type lighting because it is energy efficient but uses HSP (High Pressure Sodium) mostly which is high on energy consumption and having lesser lifespan and more management costs as compared to Light Emitting Diode.

# Data Preparation:

## Select Data:

**Relevant attributes are selected for analysis, focusing on factors potentially influencing type of light. For this I used the SLECT ATTRIBUTE operator: A screenshot of a computer program

Description automatically generated**

Here's why each attribute might be chosen for modelling:

Included:

1. **XY Coordinates Attribute**: XY coordinates are represented as two attributes, shows the geometry of the shape.
2. **Object ID**: Object ID serves as a unique identifier for each street light in the dataset. While it may not directly contribute to the classification of light types, it helps in data management and tracking individual street lights.
3. **Head Style**: The head style of street lights refers to the design or shape of the light fixture. Different head styles may correspond to different types of street lights, as certain designs are more suitable for specific applications or environments.
4. **Refractor**: The refractor attribute indicates the type of refractor or lens used in the street light fixture. The choice of refractor can affect the distribution and intensity of light emitted, which may vary for different types of street lights.
5. **Light Type:** This is the target attribute for classification, indicating the type of street light (e.g., street light, pedestrian light, lawn light). It is the attribute we want to predict based on other features.
6. **Bracket Length**: Bracket length refers to the length of the bracket supporting the street light fixture. Different types of street lights may require brackets of varying lengths depending on their design and installation requirements.
7. **Pole Height**: Pole height represents the height of the pole supporting the street light fixture. Pole height can influence the coverage area and illumination pattern of street lights, which may be indicative of different types of lighting infrastructure.
8. **Pole Type:** Pole type denotes the material or construction of the pole supporting the street light fixture. Certain pole types may be more commonly associated with specific types of street lights due to factors like durability, cost, or aesthetic considerations.
9. **Number of Lights**: The number of lights attribute indicates the total number of light sources within a street light fixture. Different types of street lights may have varying numbers of lights based on their intended purpose and lighting requirements.
10. **Wattage:** Wattage refers to the power consumption or energy output of the street light fixture. It can provide insights into the brightness and intensity of light emitted, which may differ for different types of street lights.
11. **Light Source**: The light source attribute indicates the type of technology used to produce light in the street light fixture, such as LED (Light Emitting Diode), HPS (High Pressure Sodium), LPS or others. Different light sources have distinct characteristics in terms of energy efficiency, color temperature, brightness, and longevity. These characteristics can significantly impact the appearance and performance of the street light, making the light source attribute essential for accurately classifying different types of street lights.
12. **Street Name**: The street name is imperative attribute of our project as this shows which area is having the greatest number of lights and which has less, also useful for determining the light type mostly being used in the which areas of Ottawa as Ottawa’s aim is to fully replaced lights to LED which this data can show us that how much we have achieved our aim for different regions.

**Excluded:**

1. **Created Date and Modified Date:** These attributes represent when the data entry was created or last modified. For classifying light types, the timing of data entry isn't relevant. Light types are determined by physical characteristics and specifications, not when they were recorded in the dataset.
2. **Energy Jurisdiction:** This attribute likely refers to the jurisdiction or administrative area responsible for managing the energy supply to the street lights. While important for administrative purposes, it doesn't provide information about the physical attributes of the lights themselves that would help classify them.
3. **Year 1 Dimming Wattage:** This attribute indicates the wattage of the street lights after one year of dimming. While dimming wattage might be useful for energy efficiency analysis or maintenance planning, it doesn't directly relate to the classification of light types.
4. **Light Installation Date and Pole Installation Date:** These attributes denote when the lights and poles were installed. Similar to the created and modified dates, installation dates are administrative details that don't directly influence the classification of light types.
5. **Pole ID:** Pole ID serves as a unique identifier for each pole. While essential for tracking and maintenance purposes, it doesn't provide information about the type of light mounted on the pole.

## Clean Data:

Missing values are present in Light type and refractor attribute as mentioned previously in verify data phase of data understanding. For this I have used the **Replace Missing Values** Operator: A screenshot of a computer

Description automatically generatedA diagram of a missing values

Description automatically generated

For Light type and Refractor attributes the required value is SL which is street light and drop glass respectively used as mean calculation to replace missing values.

The noise which we discussed about in the verify data quality step is not noise, but it is the different types of head styles present according to the pole height, light type, bracket length etc., So it is needed as it is already present in the dataset.

## Construct Data:

Derivation of attributes upon conversion of attributes, when the required attributes get converted to numerical or normalized for classification purposes the attributes with polynomial datatype other than class label will converted to binary types representing different columns in the outcome.

For example:

The attributes named pole type, head style and light source are converted into numeric data types for classification purposes whereas other attributes stay same as well as class label attribute which in this case is Light type.

**Data Sampling**: As the dataset is large, considered applying data sampling techniques to reduce its size while preserving the integrity of the data. A filter example with a filter

Description automatically generated with medium confidence This can improve computational efficiency without sacrificing the representativeness of the dataset. In here we reduced the data to 10,186 instances from 70000 plus instances, on the basis of several street names and converted them to the regions again for more efficiency using Filtered Examples operator with these street names with their corresponding area names : *BANK STREET - Downtown, BRONSON AVENUE - Centretown, KENT STREET - Centretown, 10TH LINE ROAD - Orleans, SPARKS - Downtown, SUSSEX DRIVE - Lower Town, QUEEN STREET - Centretown, BASELINE ROAD - Nepean, HUNT CLUB ROAD - Hunt Club, CARLING AVENUE - Carlingwood, PRINCE OF WALES DRIVE - Meadowlands, MARIVALE ROAD - Nepean, WOODROFFE AVENUE - Nepean, RIDEAU STREET - Downtown, RICHMOND ROAD - Westboro, INNES ROAD - Orleans, WELLINGTON STREET - Downtown, RIVERDALE AVENUE - Westboro, WALKLEY ROAD - South Keys, MONTREAL ROAD - Vanier, RIVERSIDE DRIVE - Riverside South, INDUSTRIAL AVENUE - Trainyards, COLONEL BY DRIVE - Sandy Hill, ALTA VISTA DRIVE - Alta Vista, ST. JOSEPH BOULEVARD - Orleans, PINECREST DRIVE - Pinecrest, GREENBANK ROAD - Barrhaven, ROBERTSON ROAD - Bells Corners, GREENFIELD AVENUE - Centrepointe, ELGIN STREET - Downtown, PRESTON STREET - Little Italy, BEECHWOOD AVENUE - New Edinburgh, SOMERSET STREET WEST - Chinatown, BAYSHORE DRIVE - Bayshore, CARLETON AVENUE - Carleton Heights, OLD RICHMOND ROAD - Richmond, OGILVIE ROAD - Beacon Hill*.

And used the Generate Attribiutes operator for genrating new attribute names REGION by the use of expression shown in picture below. A close-up of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

## Integrate Data:

Data integration techniques are applied to merge data obtained after outlier detection using both the LOF and ISF where we joined the original columns of data and outlier columns from both the methods and outlier flag column.

## Format Data:

Data formatting ensures consistency in attribute types and values, preparing the dataset for modeling. In here we removed required attributes to format data in expected form.

For example, we removed the Street name, created date and modified date, pole and light installation date, energy jurisdiction, pole id, and 1 year dimming as these are not relevant and considered as redundant, for the classification purposes.

**Data Type Adjustment:** Changed the data types for the head style, refractor, pole type, light source, Region attributes for applying the distance-based methods such as KNN and Decision Tree using the Nominal to Numerical Operator. A diagram of a mathematical equation

Description automatically generated A screenshot of a computer

Description automatically generated

**Normalization:** After that I normalized the data**A computer screen shot of a square

Description automatically generated** for X and Y shape geometry so that these attributes should not make unusually large or small impact on the output of the data.

# Modelling:

In modelling I am using the ISF and LOF for outlier detection and for classification, KNN and Decision Tree and Random Forest.

Outlier Detection:

ISF: Isolation Forest outlier detection means isolating anomalies in the dataset by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. For example, here I am using the ISF operator in RapidMiner with parameters mentioned below in picture,

A green square with orange and green text

Description automatically generated A screenshot of a computer

Description automatically generated

and this is selecting random attribute, and you can see the minimum and maximum values of that feature.

LOF: Local Outlier Factor is another method for outlier detection it works by identifying the outliers based on deviation from the local density of their neighbors. It measures the local density of a data point with respect to its neighbors and compares it to the densities of its neighbors. Points with significantly lower densities compared to their neighbors are considered outliers. LOF can capture outliers that are within dense regions of the data, making it effective for detecting outliers in clusters or groups. For example, in my project I used the LOF operator for this type of detection, with these parameters.

A computer screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

And this is the image joined output of both the methods. A screenshot of a computer

Description automatically generated

## Classification:

### KNN: K- Nearest Neighbors simple and intuitive algorithm used for both classification and regression tasks, the prediction for a new data point is determined by the majority class (for classification) or the average value (for regression) of its k nearest neighbors in the feature space, KNN's performance can be sensitive to the choice of the distance metric and the value of k.

For example, In my project for Street Lights I am using the Light Type attribute as a class attribute and used the KNN and split data usingA diagram of a diagram

Description automatically generated with medium confidence

Split Data operator for dividing data into 70/30 percentage

and A diagram of a cross-validation process

Description automatically generatedalso using the performance split, cross validation and apply model operator for testing the efficiency of KNN on different k values on testing data. Here first I applied KNN, used K as 5, 10, 13 and 20 andA screenshot of a computer

Description automatically generated then applied the model for testing the genraklization of model on untrained data acnd for checking the accuracy on each K value with R squared values and decision metrics.

**For k =5**

**Confusion Matrix** A white rectangular object with a group of people

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

**For k= 10**

**Confusion Matrix** A white grid with many squares

Description automatically generated with medium confidence

**A screenshot of a computer

Description automatically generated**

**For k= 13**

**Confusion Matrix A white rectangular object with black text

Description automatically generated with medium confidence**

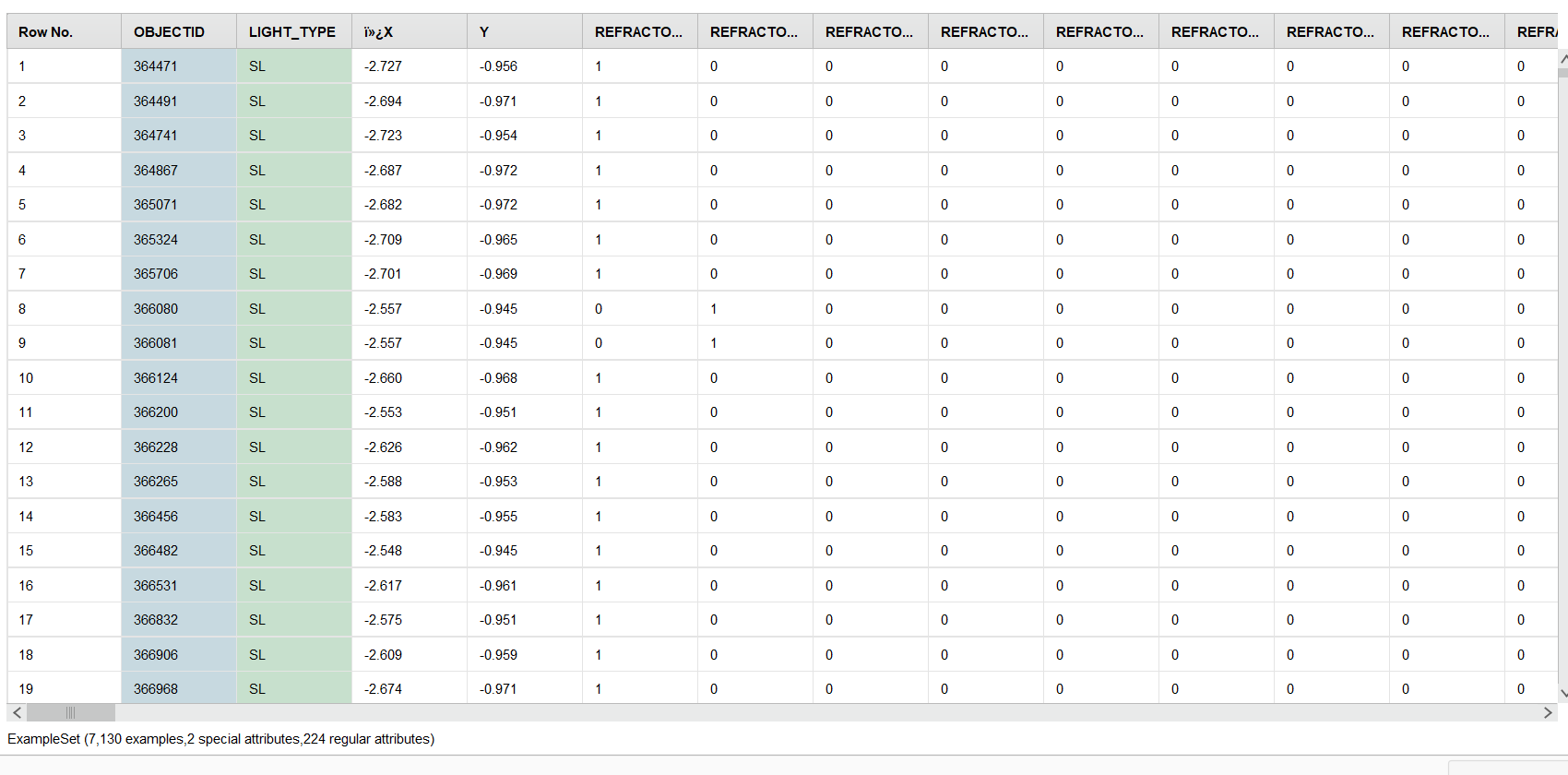
**A screenshot of a computer

Description automatically generated**

**For k= 20**

**Confusion MatrixA screenshot of a computer screen

Description automatically generated**



Here are the results for K value of 5, 10, 13 and 20 with the accuracies of 99.57, 99.41, 99.15 and, 98.66 root mean squared value as 0.050, 0.061, 0.069 and 0.089.

|  |  |  |  |
| --- | --- | --- | --- |
| K values | Accuracy | Classification error | Root Mean Squared |
| 5 | 99.57% | 0.43% | 0.050 |
| 10 | 99.41% | 0.59% | 0.61 |
| 13 | 99.15% | 0.85% | 0.069 |
| 20 | 98.66% | 1.34% | 0.89 |

### Decision Tree:

For Decision Tree I used the Operator Decision Tree, Performance Split, Cross Validation and Apply Model for evaluation of model. For classification I used these parameters maximal depth- 10, 15, 20, 50, applied pruned and unpruned for each type, with minimal leaf size 2, 5, 7 and 8 A screenshot of a computer

Description automatically generatedwith minimal size of split as 4. A diagram of a program

Description automatically generated

**Parameters:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Depth | Pruning | Leaf size | Leaf split | Confidence |
| 10 | Both | 2 | 4 | 0.1 |
| 15 | Pruned | 5 | 4 | 0.1 |
| 20 | Unpruned | 7 | 4 | 0.1 |
| 50 | Pruned | 8 | 4 | 0.1 |

**Pruned with depth of 10**

A diagram of a computer

Description automatically generated

**Confusion Matrix:** **A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

**Unpruned with depth of 10**

A diagram of a network

Description automatically generated with medium confidence

**Confusion Matrix:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Pruned with depth of 15**

A diagram of a flowchart

Description automatically generated

**Confusion Matrix:** **A white rectangular object with many small black and white objects

Description automatically generated with medium confidence**

A screenshot of a computer

Description automatically generated

**Unpruned with depth of 20**

A diagram of a company

Description automatically generated with medium confidence

**Confusion Matrix:** **A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

**Pruned with depth of 50**

A diagram of a computer

Description automatically generated with medium confidence

**Confusion Matrix:**

A white grid with many different colored lines

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Depth | Accuracy | Classification Error | Root Mean Squared |
| 10 (pruned) | 99.05% | 0.95% | 0.094 |
| 10 (Unpruned) | 99.08% | 0.92% | 0.092 |
| 15 (Pruned) | 99.02% | 0.98% | 0.095 |
| 20 (Unpruned) | 99.02% | 0.98% | 0.095 |
| 50 (Pruned) | 99.02% | 0.98% | 0.095 |

These are results for accuracy, classification error and Root mean squared values for Decision Tree with confusion matrices and screenshots of model being tested on 30% data.

## Random Forest:

Random Forest is an ensemble learning technique used in machine learning for both classification and regression tasks. It operates by constructing a multitude of decision trees during the training phase and outputs the mode (for classification) or the mean prediction (for regression) of the individual trees. In here I used the Random Forest Operator, Performance Split, Cross Validation and Apply Model for evaluation of model.

A diagram of a forest

Description automatically generated

A screenshot of a computer

Description automatically generated

**With 5 Trees : A diagram of a network

Description automatically generated with medium confidence**

**A white rectangle with black text

Description automatically generated**

**A computer screen shot of a computer

Description automatically generated**

**A diagram of a computer

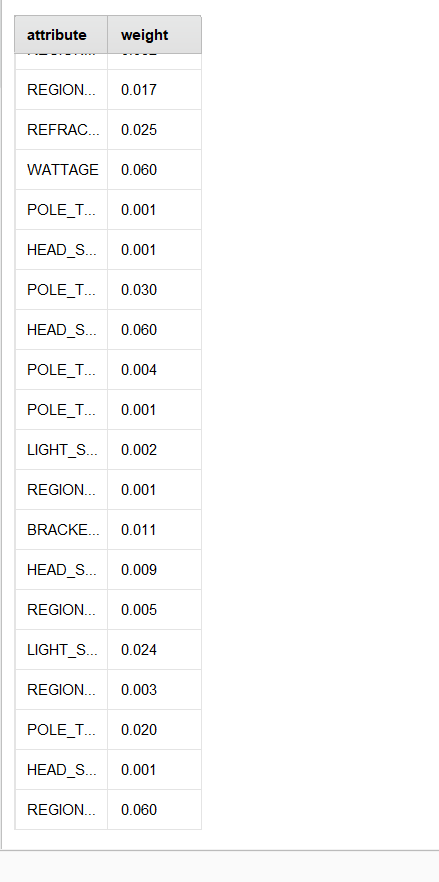
Description automatically generated with medium confidence**

**Weights with each attribute:** A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



**Confusion Matrix:** A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Accuracy: 90.74%, Classification Error: 9.26% and Root Mean Squared Eror: 0.244**

**Parameters as Number of Trees-5, Maximal depth-10 and unpruned.**

# Conclusion:

In conclusion from all the three methods KNN performed

**For KNN:**

|  |  |  |  |
| --- | --- | --- | --- |
| K values | Accuracy | Classification error | Root Mean Squared |
| 5 | 99.57% | 0.43% | 0.050 |
| 10 | 99.41% | 0.59% | 0.61 |
| 13 | 99.15% | 0.85% | 0.069 |
| 20 | 98.66% | 1.34% | 0.89 |
|  |  |  |  |

very well on testing data in all the given parameters with highest accuracy of 99.66% with value of K as 20. Whereas Decision Tree performed low as compared to KNN but still better than Random Forest with the highest accuracy of 99.08% with the given parameters of 50 depth with pruning.

**For Decision Tree:**

|  |  |  |  |
| --- | --- | --- | --- |
| Depth | Accuracy | Classification Error | Root Mean Squared |
| 10 (pruned) | 99.05% | 0.95% | 0.094 |
| 10 (Unpruned) | 99.08% | 0.92% | 0.092 |
| 15 (Pruned) | 99.02% | 0.98% | 0.095 |
| 20 (Unpruned) | 99.02% | 0.98% | 0.095 |
| 50 (Pruned) | 99.02% | 0.98% | 0.095 |

At the End Random Forest performed less efficient than KNN and Decision Tree with the highest accuracy of 90.74% on uprunning and maximum number of trees as 5.

So, the KNN is the Best amongst the all the modelling techniques applied over the course of this project and showed great efficiency not only on trained set but most efficiently on Untrained set as well.

References:

[1]

“Street Lights,” *open.ottawa.ca*. https://open.ottawa.ca/datasets/street-lights (accessed Mar. 29, 2024).

‌

‌