A black and white map pointers

Description automatically generated Trips by

Distance

St. Clair College, Mississauga

**Supervisor:**

Savita Seharawat

**Presenters:**

Gurtaranjit kaur (0814943)

Ankita (0820948)

Nimrat Singh (0810057)

**Table of Contents**

|  |  |
| --- | --- |
| Abstract | 1 |
| Keywords | 2 |
| Research Question | 2 |
| Tools | 2 |
| GitHub Source | 2 |
| Introduction | 3 |
| Literature Review | 3 |
| Methodology | 5 |
| Data Details | 7 |
| Data Preprocessing | 7 |
| Merging Data | 10 |
| Detailed Data Dictionary | 11 |
| Missing Value | 12 |
| Criteria for cleaning missing value | 13 |
| Exploratory data analyse | 15 |
| Correlation | 19 |
| Changing data types | 21 |
| Dropping columns | 22 |
| Min-Max Scalers | 25 |
| Experimental Designs | 26 |
| KNeighbors Classifier | 30 |
| DecisionTree Classifier | 32 |
| Accuracy | 33 |

**Abstract**

This dataset impressed us a lot. We got it from catalog.data.gov. The purpose of utilizing this information is to identify people who travel away from home or stay at home due to their mobility. Through our new mobility data, the Bureau of Transportation data (BTS) can now answer such queries. The travel statistics are derived from a nationwide panel of mobile device data gathered from various sources.

Data analysis is carried out at the national, state, and county levels. A weighting process broadens the sample of millions of mobile devices, ensuring that the findings are representative of the total population of a country, state, or county.

Trips are moves that include a stay of more than 10 minutes at an unidentified site. Furthermore, we investigate the effect of distance on mode choice, demonstrating how consumers prefer different modes of transportation dependent on journey distance. Longer travels are more likely to utilize private vehicles, trains, buses, or airplanes, whereas shorter journeys are more likely to involve walking, cycling, or local public transit.

Understanding these travel patterns and how they affect mobility has several practical applications. This data may be used by urban planners to create more efficient transport networks, prioritize infrastructure development, and adopt policies that encourage sustainable and ecologically friendly travel choices. Furthermore, transport and tourist enterprises can customize their services to individual travel demands depending on distance classifications.

Finally, this study sheds light on the complicated link between travel behavior and distance, emphasizing the significance of considering varied travel patterns while developing mobility plans. Stakeholders may collaborate to create a more accessible, efficient, and resilient transportation landscape for the future by recognizing the specific features of journeys by distance.

For this project, we are utilizing predictive analysis as a theme and Python as a tool.

**Keywords:** Train\_test model, DecisionTreeClassifier, Min\_Max Scaler, confusion matrix.

**Research Questions:**

1. **What are some factors to consider when planning a trip of long distance?**
2. **How does the distance of a trip influence the budget and cost of travel?**
3. **Which states and counties have more people taking trips?**
4. **How does the length of the trip impact the amount of planning and research required?**
5. **How many people are active for trip from total population?**

**Tools:**

All formulation and data visualization will be done by using Python.

**Source:**

<https://catalog.data.gov/dataset/trips-by-distance>

Introduction:

The Maryland Transportation Institute and the Centre for Advanced Transportation Technology Laboratory at the University of Maryland estimate the Trips by Distance statistics and the number of individuals remaining and not staying at home for the Bureau of Transportation Statistics. Data analysis is carried out at the national, state, and county levels. A weighting process broadens the sample of millions of mobile devices, ensuring that the findings are representative of the total population of a country, state, or county. No data for a county is supplied to ensure confidentiality and data integrity.

To ensure anonymity and data quality, no data for a county is published if there are less than 50 devices in the sample on any given day.

Trips are moves that include a stay of more than 10 minutes in an unidentified area away from home. Home addresses are assigned on a weekly basis. Numerous excursions are defined as movements with numerous stays of more than 10 minutes before returning home. Trips include all means of transportation, including automobile, train, transit, and air travel.

**Literature review**

An examination of the literature on the issue of "trip by distance" finds a corpus of research focusing on the link between travel patterns and distance travelled. Several studies have been conducted to investigate the effects of distance on trip features, mode choice, and transportation behavior.

Researchers investigated how the distance of a journey influences mode selection, with data indicating that shorter distances generally support walking or cycling, whilst longer ones may lead to a preference for motorized transportation.

Furthermore, studies examining the impact of journey distance on travel time and energy consumption have been conducted. Longer travels are associated with greater energy costs and longer travel times.

The literature also emphasizes the significance of urban design and transit infrastructure in affecting trip distances. Shorter travel lengths and less reliance on private automobiles have been linked to well-connected public transport networks and diverse land-use patterns.

Furthermore, research on the psychological elements of journey distance provides information on how perceptions of distance impact travel decisions and participation in activities.

Because of its potential to optimize routes, reduce fuel consumption, and increase overall efficiency, this method of trip planning and analysis has received a lot of attention in recent years. This overview of the literature delves into the important studies and research on the notion of "Trip by Distance" and its applications in numerous industries.

The transportation and logistics sector has been among the first to see "Trip by Distance" as a helpful statistic for optimizing delivery routes and lowering costs. Kapoor et al. (2018) illustrated how adding real-time GPS data might improve performance.

Kapoor et al. (2018) demonstrated how combining real-time GPS data with trip distance estimations increased a large logistics company's delivery performance by 15%. The study emphasized the significance of precise distance measuring in lowering fuel consumption and emissions, resulting in considerable environmental advantages.

"Trip by Distance" has shown potential in urban planning and traffic management, easing congestion, and enhancing overall traffic flow. Liang et al. (2019) examined journey patterns in a metropolitan region using GPS data from ride-sharing services. The researchers were able to recommend more effective transport methods by emphasizing distance-based route planning, resulting in shorter travel times and smoother traffic movement.

"Trip by Distance" has been investigated in the context of public transport as a technique of encouraging sustainable travel behavior. Wu et al. (2020) evaluated the viability of distance-based pricing for public transport networks. According to the findings of the study, establishing such a system not only increased the use of public transport but also decreased the financial burden on low-income passengers, making it a fairer method.

The environmental advantages of "Trip by Distance" have received much attention in the literature. Greenberg et al. (2017) examined the carbon footprint of several modes of transport and discovered that distance-based mobility, such as cycling and walking, considerably decreased emissions when compared to time-based options such as driving or taking public transport. These findings highlight the potential of "Trip by Distance" as a tool to encourage environmentally friendly travel decisions.

"Trip by Distance" has been connected to increased physical activity and well-being in the context of health. Andersen et al. (2019) investigated the health effects of distance-based commuting techniques, discovering that those who used active transportation to go to work, such as walking or cycling, reported greater levels of physical activity and overall health satisfaction.

The "Trip by Distance" idea has evolved as a helpful technique in a variety of fields, including transportation and logistics, urban planning, public transit, and sustainability. The assessment of the literature emphasized the significance of precise distance measuring in optimizing travel routes, lowering environmental consequences, encouraging sustainable transportation options, and enhancing safety.

Trip by Distance is an important idea that has gained traction because of its implications in a variety of fields including urban planning, transportation management, logistics, tourism, and environmental conservation. Measuring journeys by distance enables academics, policymakers, and stakeholders to examine and analyze mobility patterns and resource usage more effectively.

Trips by Distance are important in urban planning since it helps to build efficient transportation networks and optimize land use. According to research, towns with well-planned and accessible public transportation systems that emphasize shorter journey durations have less traffic congestion and better air quality. Furthermore, compact urban development concepts that allow for reduced commute distances have been linked to healthier and more lively communities.

To improve efficiency and safety, transport systems mainly rely on the notion of Trip by Distance. The application of this statistic in transportation management assists in the optimization of route planning, scheduling, and resource allocation, resulting in cost savings and higher service quality. Furthermore, precise trip distance data is critical for the development of trip-planning applications and navigational systems.

Trip by Distance is useful in logistics and supply chain management for simplifying distribution routes and lowering total transportation expenses. Companies may increase their environmental sustainability while simultaneously assuring timely delivery of goods by minimizing distances travelled throughout the delivery process.

Trip by Distance is an important component in the tourism sector, as travelers frequently prioritize short-distance journeys. To improve efficiency and safety, transport systems mainly rely on the notion of Trip by Distance. The application of this statistic in transportation management assists in the optimization of route planning, scheduling, and resource allocation, resulting in cost savings and higher service quality. Furthermore, precise trip distance data is critical for the development of trip-planning applications and navigational systems.

Trip by Distance is useful in logistics and supply chain management for simplifying distribution routes and lowering total transportation expenses. Companies may increase their environmental sustainability while simultaneously assuring timely delivery of goods by minimizing distances travelled throughout the delivery process.

Trip by Distance is an important component in the tourism sector, as travelers frequently prioritize short-distance journeys. However, trip by distance is a strong statistic, it does have certain drawbacks. Accurate data collecting can be challenging, especially in rural or distant places with little infrastructure. Furthermore, this statistic may not represent the full complexities of travel behavior, such as the motivations for long-distance travel or changes in transportation options. The introduction of technology, such as GPS tracking, smart sensors, and data analytics, has revolutionized trip distance data gathering and analysis. These developments allow for real-time monitoring and informed decision-making, opening new avenues for enhancing transportation systems and urban planning.

Trip by Distance has evolved as an important tool for studying and controlling travel patterns in a variety of industries. Trip by Distance has evolved as an important tool for studying and controlling travel patterns in a variety of industries. Its use in urban planning, transportation management, logistics, tourism, and environmental conservation demonstrates its adaptability and societal influence. To fully realize the concept's promise, researchers and policymakers must overcome the hurdles and invest in technical breakthroughs that allow for more accurate and extensive data collecting. Trips by Distance can lead to more sustainable and efficient transportation systems that benefit both people and the environment.

Historical Background and Conceptual Framework: The notion of calculating trip distances dates to early transportation design. Initially, it was utilized to improve the efficiency of transport networks and properly distribute resources. With advances in technology and data analytics, the use of Trip by Distance has grown to include a broader variety of fields like urban planning, logistics, tourism, and environmental protection.

Trip by Distance in Urban Planning: Trip by Distance is a key statistic in urban planning for creating efficient transport networks and guiding urban growth. According to studies, towns with lower average commute lengths have less traffic congestion, better accessibility, and better air quality. Urban regions can support economic development by including Trip by Distance into the planning process.

Urban regions may develop sustainable, walkable communities that promote physical activity and social connections by including Trip by Distance in the design process.

Transportation Management and Optimization: Accurate assessment of trips by distance is critical in transportation management for optimizing routes, scheduling, and resource allocation. This statistic aids in the development of intelligent transportation systems that allow for real-time monitoring and data-driven decision-making. Governments and businesses may cut operating costs, enhance service quality, and lessen the environmental effect of transportation by implementing Trip by Distance in transportation management.

Trip by Distance in Logistics and Supply Chain Management: Using Trip by Distance in logistics and supply chain management provides for more effective distribution networks. Companies can cut fuel usage by reducing the distances travelled during the transportation of goods. As a result, environmental sustainability and supply chain resilience are improved.

Tourists and Travel Behavior: In the tourist sector, trip by distance have a considerable impact on travel behavior. Shorter excursions are frequently prioritized by travelers to maximize their time at locations while minimizing travel-related expenditures. Furthermore, the notion is critical in encouraging sustainable travel practices such as eco-tourism, which emphasizes responsible travel choices that reduce negative environmental consequences.

Environmental Impact and Sustainability: Shortening average journey lengths directly improves environmental sustainability. Shorter journey distances are associated with fewer greenhouse gas emissions, lower air pollution, and lower energy usage, according to studies. As cities and governments place a greater emphasis on environmental conservation and climate change mitigation, including Trip by Distance into transportation and urban planning is becoming more common.

Challenges and Limitations: While travelling by distance has many advantages, it also has certain drawbacks. One of the most difficult issues is collecting reliable data, especially in rural or distant locations with minimal infrastructure. Furthermore, this statistic may not completely represent the intricacies of travel behavior, such as the motivations for long-distance travel or the impact of diverse socioeconomic circumstances.

Technological Advances and Data Analytics: Technological advancements like GPS tracking, smart sensors, and data analytics have transformed the collecting and processing of trip distance data. These technological advancements allow for real-time monitoring, data integration, and predictive modelling, opening new avenues for optimizing transportation networks and urban planning.

In the future, including Trip by Distance into decision-making processes will be critical in supporting sustainable and efficient transportation networks. Policymakers and academics should work together to overcome data collecting issues, invest in new technology, and create novel methods for incorporating Trip by Distance into a wide range of applications.

Trip by Distance is an important statistic in logistics and supply chain management for optimizing distribution routes and lowering transportation costs. Companies that focus on minimizing travel lengths can increase their environmental sustainability while also increasing customer satisfaction through timely and cost-effective delivery. Furthermore, advances in supply chain analytics have allowed for the incorporation of Trip by Distance into decision-making processes, resulting in increased efficiency and lower carbon footprints.

Trip by Distance is an important component in the tourism sector since travelers frequently aim to maximize their time at locations by travelling shorter distances. Sustainable tourism programs emphasize the need to reduce the environmental effect of travel-related activities while also helping local economies. Data from Trip by Distance is essential for developing eco-tourism strategies and designing sustainable travel routes.

While a trip by distance is a strong measure, its implementation is fraught with difficulties and restrictions. Accurately collecting trip distance data can be difficult, particularly in distant or underdeveloped locations with little infrastructure. Furthermore, because factors such as trip purpose and individual preferences might impact the mode of transportation and trip distance, this statistic may not completely represent the complexities of travel behavior.

Recent technical improvements have substantially enhanced trips by distance data gathering, processing, and application. GPS technology, smart sensors, and data analytics have transformed data collecting, allowing for real-time monitoring of travel patterns and dynamic traffic management. Furthermore, data-driven decision-making processes have improved, resulting in improved urban planning, transportation management, and logistics optimization.

**Methodology**

**Data Details**

Dataset contains 22 columns about people which explain data about the time how many times they stay at home or go outside from their home within 10 minutes throughout their mobility, even though their record of staying at home or going out of home is calculated weekly and monthly also. This dataset contains the counts of people which move from their home to somewhere in conditions such as- number of people <1 till 500.

**Table 1: Describing all the Categorical attributes.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Categorical Attribute** | **Description** | **Levels** | **Count** |
| Level | This attribute describes the level of the trip. | 3 | National: 20318  State: 462233  County: 566024 |
| State\_Postal\_code | It describes the postal code of the state to which the trip  was made | Unique | None |
| County\_Name | It describes the name of county belonging of people those moved | Unique | None |

**Table 2: Summarize Numerical attributes.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Numerical Attributes** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| State\_FIPS | 30.219 | 15.16822 | 1 | 18 | 29 | 45 | 56 |
| County\_FIPS | 30380.29 | 15160.11 | 1001 | 18177 | 29175 | 45081 | 56045 |
| Population\_Staying\_at\_Home | 146291.7 | 2225797 | 8 | 2079 | 5018 | 15950 | 1.1E+08 |
| Population\_Not\_Staying\_at\_Home | 523471.4 | 7649900 | 87 | 9793 | 23474 | 71318 | 2.74E+08 |
| Number\_of\_Trips | 2477794 | 36181617 | 220 | 49784 | 120876 | 360330 | 1.57E+09 |
| Number\_o\_ Trips <1 | 612079.8 | 9003866 | 0 | 10486 | 25642 | 77753 | 4.23E+08 |
| Number\_of\_Trips\_1-3 | 619197 | 9022217 | 0 | 11384 | 30658 | 96197 | 4.05E+08 |
| Number\_of\_Trips\_3-5 | 304089.4 | 4437764 | 0 | 4794 | 13923 | 46219 | 1.98E+08 |
| Number\_of\_Trip\_5-10 | 384310 | 5615742 | 0 | 6463 | 16999 | 54469 | 2.53E+08 |
| Number\_of\_Trips\_10-25 | 375938 | 5502116 | 0 | 8001 | 19345 | 55151 | 2.57E+08 |
| Number\_of\_Trips\_25-50 | 120812.8 | 1765528 | 0 | 3860 | 8387 | 21096 | 76367324 |
| Number\_of\_Trips\_50-100 | 38847.12 | 575391.8 | 0 | 1240 | 2717 | 6679 | 28735606 |
| Number\_of\_Trips\_100-250 | 15824.53 | 240432.3 | 0 | 363 | 857 | 2372 | 14476977 |
| Number\_of\_Trips\_250-500 | 3553.21 | 53461.22 | 0 | 58 | 173 | 533 | 3651375 |
| Number\_of\_Trips >=500 | 3142.248 | 50697.72 | 0 | 27 | 98 | 426 | 5003062 |
| Week | 21.71479 | 12.42963 | 0 | 11 | 23 | 32 | 52 |
| Month | 5.575419 | 2.847674 | 1 | 3 | 6 | 8 | 12 |

**Datetime part:**

DateTime is a key notion used to describe points in time in computer science and software development. It integrates date and time information into a single data structure, making temporal data handling and comparison simple. DateTime is a programming language and framework that is commonly used to handle time-related tasks such as event scheduling, time zone conversions, and date arithmetic.

|  |  |  |
| --- | --- | --- |
| Attributes | Starting Date | End Date |
| DateTime | 2019-01-01 | 2019-11-01 |

**Data Preprocessing:**

**Detailed Data Dictionary:**

In this step, I took one variable at one time to do analysis. First, I assigned the correct data type of each attribute. For numeric attribute I checked five summaries i.e., minimum, maximum, mean, first and third quartile. Next, I decided level and frequency table for each categorical attribute. Then I decided to remove the duplicates, missing values from my dataset.

**Missing Values:**

|  |  |
| --- | --- |
| **Attributes Name** | **No of Missing Values** |
| **State\_FIPS** | 901 |
| **State\_Postal\_Code** | 901 |
| **County\_FIPS** | 46852 |
| **County\_Name** | 46852 |
| **Population\_Staying\_at\_Home** | 12950 |
| **Population\_Not\_Staying\_at\_Home** | 12950 |
| **Number\_of\_Trips** | 12950 |
| **Number\_of\_ Trips<1** | 12950 |
| **Number\_of\_Trip\_5-10** | 12950 |
| **Number\_of\_Trips (10-25)** | 12950 |
| **Number\_of\_Trips\_25-50** | 12950 |
| **Number\_of\_Trips\_50-100** | 12950 |
| **Number\_of\_Trips\_100-250** | 12950 |
| **Number\_of\_Trips\_250-500** | 12950 |
| **Number\_of\_Trips >=500** | 12950 |

We can above mentioned attributes have more than 1000 missing values.

**Criteria for cleaning missing values:**

Next step to analyse if there are any missing values firstly remove or replace them to get accurate data. I used to mean to replace all the null values from my dataset.

**Exploratory Data Analysis:**

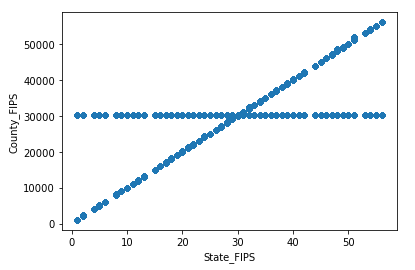
****

Figure1: Scatterplot for attributes County\_FIPS and State\_FIPS

As we can see, on y-axis is County\_FIPS and on x-axis is State\_FIPS which represent the counting of people those move out of their home or stay at home with the gap of 10 minutes.

**Histogram of State\_FIPS:**

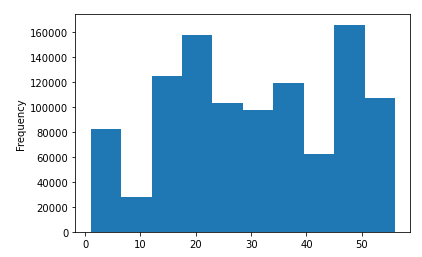
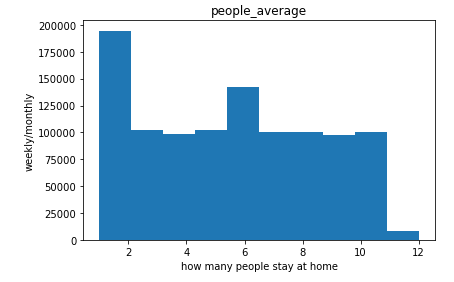


Figure 2: Histogram of State\_FIPS

In this histogram, the fluctuation is shown according to trips covered by people within fixed time on weekly and monthly by males and females even included old agers.

**Histogram of how many people stay at home:**



In this graph, I can see the plot that calculate weekly/monthly about the people who stay at home within fixed time. Even in this graph x-label is how many people stay at home and y-label is weekly/monthly.

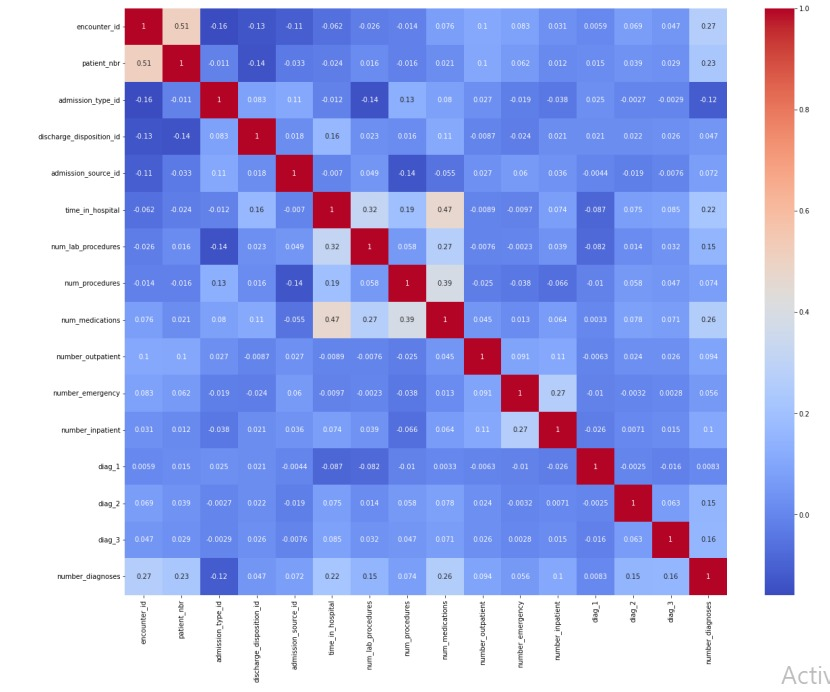
**Scatter plot of Number\_of\_Trips and Number\_of\_ Trips<1**

**A graph of blue dots

Description automatically generated**

In this graph, I can see the plot that calculate weekly/monthly about the people who stay at home within condition like number of trips by people which is less than 1 within fixed time. Even in this graph x-label is how many people stay at home and y-label is weekly/monthly.

**Correlation:**



A correlation heatmap is a graphical representation used to visualize the correlation between different variables in a dataset. In the context of "trip by distance," it would typically involve examining the relationship between trip-related variables and the distance covered during those trips. There are some variables:

**Level:** In this column, we have 3 attributes – country, national and state

**Date:** In date part, data gather according to data

**State\_FIPS:** State’s first two letters are taken in data frame.

**County\_FIPS:** This is numerical attributes in our dataset.

**Population\_staying\_at\_home:** How many people are staying at home.

**Population\_staying\_not\_at\_home:** How many people don’t stay at home.

**Number\_of\_Trips:** How many trips are covered by one person.

**Week:** Every week trips are count of people.

**Month:** After counting weekly trips, monthly trips are captured in our dataset.

**Data types from object to categorical:**

Some column has object datatypes so basically, we change their datatypes to categorical to make our dataset accurate. We change the datatypes of

Level column due to responsible for we have three attributes in under this column: national, county and state.

**Min-Max Scaling:**

It is a data normalization approach that scales data between 0 and 1. It is a simple technique to normalize data using Python’s min-max functions.

Normalization is only deal with numerical data, so I decided all category

Variables.

**Experimental Design:**

**Cross validation:**

A technique for testing efficiency of machine learning models are cross validation. For modeling, I implemented some models validation techniques. As stated below:

**Train and Test Split Approach:**

In this model, the entire data is randomly divided into training and test set. I divided the information into two parts (training and testing sets). The training set contains 80% of the records in the dataset whereas, the test set contains 20% of the dataset observation.

**KNeighborsClassifier:**

When given a new input data point, the KNN classifier identifies the k nearest data points (neighbors) to the input in the feature space. The class of the input data point is then determined by the majority class among these k neighbors. In other words, if most of the k neighbors belong to a particular class, the input data point is classified into that class.

The choice of the 'k' value is essential in KNN. A smaller 'k' might lead to noisy decisions, while a larger 'k' may cause the boundaries between classes to become too smooth. The optimal value of 'k' depends on the specific dataset and problem at hand and is often found through experimentation or cross-validation.

**DecisionTreeClassifier:**

Decision trees can be prone to overfitting, where they memorize the training data and perform poorly on new, unseen data. To mitigate this, you can tune hyperparameters like max\_depth, min\_samples\_split, or min\_samples\_leaf to control the size and complexity of the tree and improve its generalization ability.

The DecisionTreeClassifier is a powerful and interpretable classification algorithm, making it a popular choice for various machine learning tasks. However, in more complex scenarios, ensemble methods like Random Forests or Gradient Boosting might be preferred, as they can improve performance and reduce overfitting compared to individual decision trees.

**Accuracy:**

|  |  |
| --- | --- |
| Model | Accuracy |
| DecisionTreeClaasifier | 1.0 |