Capstone Project I

Sahil Singh 0817659

Jarmandeep Singh 0821567

Jasreet Kaur 0817638

In Supervision of Savita Seharawat

Date of submission: 23-05-2023

St. Clair College Ace Acumen Academy

S23 DAB322 DAB3A CAPSTONE PROJECT I

Table of Contents

**Abstract3**

Keywords3

Research Questions3

Tools4

Github Source4

**Introduction:5**

Purpose5

**Methodology6**

**Data details6**

Table 1: Summary of Categorical Attribute7

Table 2: Summary of Numeric Attribute9

Table 1: Summary of Data Attribute11

**Data Preprocessing13**

Importing Dataset13

Dealing with Duplicates and Null values13

**Descriptive Visualizations16**

**References21**

Abstract

The data for the project is sourced from data.gov. The data is managed and collected by National Center for Education Statistics' (NCES) Education Demographic and Geographic Estimate (EDGE). NCES and the EDGE programme yearly update point locations (latitude and longitude) for public elementary and secondary schools that are part of the NCES Common Core of Data (CCD). The CCD programme collects administrative and financial data on all public schools, school districts, and state education authorities in the United States on an annual basis. The information is provided by state education agency authorities and includes basic directory and contact information for schools and school districts, as well as student demographics, the number of teachers, the school grade span, and numerous other administrative conditions. The CCD school and agency point locations are calculated.

Keywords: Classification, Regression, Decision Tree, demographics, Geographical patterns

**Questions**

**Q1:** Is there evidence of spatial patterns or clusters of schools based on geographic location?

**Q2:** Project the anticipated demand for various types of educational institutions (e.g., public, private, specialized) across different geographical regions?

**Q3:** Develop a predictive model for estimating student enrollment across various grade levels in a specific school by leveraging historical enrollment patterns and demographic information?

**Q4:** Can a predictive model be developed to estimate the student-teacher ratio in schools by utilizing school characteristics and enrollment data?

**Q5:** Utilize historical patterns and location information to forecast the probability of schools undergoing administrative changes, such as modifications in grade spans or transitions between school types (e.g., elementary to middle school)?

**Tools:**

-Pyhton

**Github Source:**

**-**

**Introduction**

A selection of administrative attributes for public elementary and secondary schools from the 2020-2021 school year applied to point locations created by the National Center for Education Statistics' (NCES) Education Demographic and Geographic Estimates (EDGE) program.

The National Center for Education Statistics' (NCES) Education Demographic and Geographic Estimate (EDGE) program develops annually updated point locations (latitude and longitude) for public elementary and secondary schools included in the NCES Common Core of Data (CCD). The CCD program annually collects administrative and fiscal data about all public schools, school districts, and state education agencies in the United States. The data are supplied by state education agency officials and include basic directory and contact information for schools and school districts, as well as characteristics about student demographics, number of teachers, school grade span, and various other administrative conditions. CCD school and agency point locations are derived from reported information about the physical location of schools and agency administrative offices. The point locations and administrative attributes in this data layer were developed from the 2020-2021 CCD collection. For more information about NCES school point data, see: https://nces.ed.gov/programs/edge/Geographic/SchoolLocations. For more information about these CCD attributes, as well as additional attributes not included, see: https://nces.ed.gov/ccd/files.asp.

All information contained in this file is in the public domain. Data users are advised to review NCES program documentation and feature class metadata to understand the limitations and appropriate use of these data.

**Purpose:** The National Center for Education Statistics (NCES) Education Demographic and Geographic Estimates (EDGE) program develops data resources and information to help data users investigate the social and spatial context of education. School point locations (latitude/longitude values) are a key component of the NCES data collection. These data are needed to address a variety of spatially-oriented tasks and research questions. They provide information needed to construct NCES school-based surveys; they provide indicators needed to help determine program eligibility; and they provide the foundation for determining geographic associations with other types of entities. This document describes the content of geocode files developed for schools and LEAs reported in the Common Core of Data (CCD) school and agency universe. The CCD universe is a comprehensive collection of public elementary and secondary schools and the LEAs that operate and support them. The CCD provides administrative data about enrollment, staffing, and program participation for schools, LEAs, and states. State education agencies (SEAs) report these data to the U.S. Department of Education in a series of file submissions throughout the school year. The geocode files include the unique school and agency identifiers assigned by CCD, and this shared ID can be used to integrate the geocodes with the CCD directory files. Additional discussion and documentation for the CCD school and agency files is available at NCES's CCD website.

**Methodology:**

**Data Details:**

The data that is used here is ‘Public\_School\_Characteristics\_2020-21’. This data is derived from data.gov. This data includes the characteristics of approximately 100,000 schools in US with fields discussed below:

Public School Geocode File (EDGE\_GEOCODE\_PUBLICSCH\_*YYYY*) Record Layout.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Length** | **Type** | **Description** |
| NCESSCH | 12 | Number | School identification number |
| LEAID | 7 | Number | School district identification number |
| NAME | 60 | String | Name of institution |
| OPSTFIPS | 2 | String | FIPS state code for operating state |
| STREET | 50 | String | Reported location street address |
| CITY | 25 | String | Reported location city |
| STATE | 2 | String | Reported location state |
| ZIP | 5 | String | Reported location ZIP code |
| STFIP | 2 | String | State FIPS |
| CNTY | 5 | String | County FIPS |
| NMCNTY | 100 | String | County name |
| LOCALE | 2 | String | Locale code |
| LAT | 10.6 | Double | Latitude of school location |
| LON | 11.6 | Double | Longitude of school location |
| CBSA | 5 | String | Core Based Statistical Area |
| NMCBSA | 100 | String | Core Based Statistical Area name |
| CBSATYPE | 1 | String | Metropolitan or Micropolitan Statistical Area  indicator |
| CSA | 3 | String | Combined Statistical Area |
| NMCSA | 100 | String | Combined Statistical Area name |
| NECTA | 5 | String | New England City and Town Area |
| NMNECTA | 100 | String | New England City and Town Area name |
| CD | 4 | String | Congressional District |
| SLDL | 5 | String | State Legislative District - Lower |
| SLDU | 5 | String | State Legislative District - Upper |
| SCHOOLYEAR | 9 | String | School year |

Public Local Education Agency Geocode File (EDGE\_GEOCODE\_PUBLICLEA\_*YYYY*) Record Layout

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Length** | **Type** | **Description** |
| LEAID | 7 | Number | Agency identification number |
| NAME | 60 | String | Name of agency |
| OPSTFIPS | 2 | String | FIPS state code for operating state |
| STREET | 50 | String | Reported location street address |
| CITY | 25 | String | Reported location city |
| STATE | 2 | String | Reported location state |
| ZIP | 5 | String | Reported location ZIP code |
| STFIP | 2 | String | State FIPS |
| CNTY | 5 | String | County FIPS |
| NMCNTY | 100 | String | County name |
| LAT | 10.6 | Double | Latitude of agency office location |
| LON | 11.6 | Double | Longitude of agency office location |
| CBSA | 5 | String | Core Based Statistical Area |
| NMCBSA | 100 | String | Core Based Statistical Area name |
| CBSATYPE | 1 | String | Metropolitan or Micropolitan Statistical Area  indicator |
| CSA | 3 | String | Combined Statistical Area |
| NMCSA | 100 | String | Combined Statistical Area name |
| NECTA | 5 | String | New England City and Town Area |
| NMNECTA | 100 | String | New England City and Town Area name |
| CD | 4 | String | Congressional District |
| SLDL | 5 | String | State Legislative District - Lower |
| SLDU | 5 | String | State Legislative District - Upper |
| SCHOOLYEAR | 9 | String | School year |
| LOCALE | 2 | String | Locale assigned to district |
| PCT\_CITY11 | 8 | Number | Percentage of enrolled students attending schools  in locale 11 (city – large) |
| PCT\_CITY12 | 8 | Number | Percentage of enrolled students attending schools  in locale 12 (city – midsize) |
| PCT\_CITY13 | 8 | Number | Percentage of enrolled students attending schools  in locale 13 (city – small) |
| PCT\_SUB21 | 8 | Number | Percentage of enrolled students attending schools  in locale 21 (suburb – large) |
| PCT\_SUB22 | 8 | Number | Percentage of enrolled students attending schools  in locale 22 (suburb – midsize) |
| PCT\_SUB23 | 8 | Number | Percentage of enrolled students attending schools  in locale 23 (suburb – small) |
| PCT\_TOWN31 | 8 | Number | Percentage of enrolled students attending schools  in locale 31 (town – fringe) |
| PCT\_TOWN32 | 8 | Number | Percentage of enrolled students attending schools  in locale 32 (town – distant) |
| PCT\_TOWN33 | 8 | Number | Percentage of enrolled students attending schools  in locale 33 (town – remote) |
| PCT\_RURAL41 | 8 | Number | Percentage of enrolled students attending schools  in locale 41 (rural – fringe) |
| PCT\_RURAL42 | 8 | Number | Percentage of enrolled students attending schools  in locale 42 (rural – distant) |
| PCT\_RURAL43 | 8 | Number | Percentage of enrolled students attending schools  in locale 43 (rural - remote) |

|  | count | mean | std | min | 25% | 50% | 75% | max |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 100722.00 | -93.00 | 17.54 | -176.64 | -101.87 | -89.52 | -81.24 | 145.78 |
| Y | 100722.00 | 37.82 | 5.75 | -14.35 | 33.95 | 38.74 | 41.67 | 71.30 |
| OBJECTID | 100722.00 | 50361.50 | 29076.08 | 1.00 | 25181.25 | 50361.50 | 75541.75 | 100722.00 |
| NCESSCH | 100722.00 | 289465747067.07 | 165894838501.06 | 10000500870.00 | 130261004327.25 | ######## | 421899003813.25 | 780003000034.00 |
| LEAID | 100722.00 | 2894657.44 | 1658948.39 | 100005.00 | 1302610.00 | ######## | 4218990.00 | 7800030.00 |
| LZIP | 100722.00 | 54070.68 | 29046.79 | 601.00 | 30165.00 | 55305.00 | 78611.00 | 99950.00 |
| STATUS | 100722.00 | 1.08 | 0.59 | 1.00 | 1.00 | 1.00 | 1.00 | 8.00 |
| TOTFRL | 76964.00 | 280.16 | 293.88 | 3.00 | 94.00 | 206.00 | 374.00 | 10595.00 |
| FRELCH | 74117.00 | 258.41 | 272.27 | 0.00 | 81.00 | 185.00 | 349.00 | 6655.00 |
| REDLCH | 74117.00 | 24.98 | 35.88 | 0.00 | 3.00 | 15.00 | 33.00 | 1365.00 |
| PK | 31701.00 | 34.08 | 45.39 | 0.00 | 11.00 | 24.00 | 43.00 | 1956.00 |
| KG | 53537.00 | 63.35 | 45.99 | 0.00 | 36.00 | 59.00 | 83.00 | 2441.00 |
| G01 | 53941.00 | 65.60 | 44.50 | 0.00 | 38.00 | 62.00 | 87.00 | 2333.00 |
| G02 | 53980.00 | 65.68 | 44.38 | 0.00 | 38.00 | 62.00 | 87.00 | 2393.00 |
| G03 | 53932.00 | 66.17 | 45.18 | 0.00 | 38.00 | 62.00 | 88.00 | 2347.00 |
| G04 | 53722.00 | 67.46 | 47.31 | 0.00 | 38.00 | 63.00 | 90.00 | 2363.00 |
| G05 | 52426.00 | 69.88 | 54.75 | 0.00 | 37.25 | 63.00 | 91.00 | 2454.00 |
| G06 | 37220.00 | 101.16 | 108.05 | 0.00 | 26.00 | 65.00 | 136.00 | 2467.00 |
| G07 | 32150.00 | 120.65 | 129.40 | 0.00 | 23.00 | 73.00 | 188.00 | 2574.00 |
| G08 | 32378.00 | 120.69 | 131.41 | 0.00 | 22.00 | 71.00 | 188.00 | 2595.00 |
| G09 | 26986.00 | 149.40 | 194.42 | 0.00 | 11.00 | 69.00 | 215.00 | 3307.00 |
| G10 | 26847.00 | 145.72 | 189.13 | 0.00 | 13.00 | 67.00 | 205.00 | 3164.00 |
| G11 | 26819.00 | 138.50 | 178.41 | 0.00 | 15.00 | 64.00 | 192.00 | 3000.00 |
| G12 | 26731.00 | 137.13 | 173.73 | 0.00 | 18.00 | 68.00 | 191.00 | 2850.00 |
| G13 | 143.00 | 13.60 | 16.36 | 0.00 | 1.00 | 7.00 | 23.50 | 80.00 |
| UG | 8153.00 | 15.20 | 37.65 | 0.00 | 1.00 | 5.00 | 15.00 | 802.00 |
| AE | 182.00 | 26.20 | 272.42 | 0.00 | 0.00 | 0.00 | 0.00 | 3646.00 |
| TOTMENROL | 0.00 |  |  |  |  |  |  |  |
| TOTFENROL | 0.00 |  |  |  |  |  |  |  |
| TOTAL | 98651.00 | 500.98 | 475.92 | 0.00 | 233.00 | 408.00 | 624.00 | 21049.00 |
| MEMBER | 98651.00 | 500.93 | 475.83 | 0.00 | 233.00 | 408.00 | 624.00 | 21049.00 |
| FTE | 91220.00 | 33.05 | 26.18 | 0.00 | 17.47 | 28.00 | 41.35 | 1093.00 |
| STUTERATIO | 99506.00 | 14.37 | 60.80 | -2.00 | 11.00 | 14.03 | 17.10 | 17900.00 |
| AMALM | 69289.00 | 3.63 | 15.36 | 0.00 | 0.00 | 1.00 | 2.00 | 618.00 |
| AMALF | 69035.00 | 3.50 | 14.97 | 0.00 | 0.00 | 1.00 | 2.00 | 690.00 |
| AM | 76349.00 | 6.46 | 28.81 | 0.00 | 0.00 | 1.00 | 3.00 | 1308.00 |
| ASALM | 82938.00 | 16.47 | 46.50 | 0.00 | 1.00 | 3.00 | 12.00 | 2044.00 |
| ASALF | 82625.00 | 15.68 | 43.92 | 0.00 | 1.00 | 3.00 | 12.00 | 1473.00 |
| AS | 86894.00 | 30.62 | 88.20 | 0.00 | 1.00 | 5.00 | 22.00 | 3517.00 |
| BLALM | 89123.00 | 42.14 | 76.86 | 0.00 | 2.00 | 11.00 | 48.00 | 2619.00 |
| BLALF | 88434.00 | 40.93 | 76.46 | 0.00 | 2.00 | 10.00 | 47.00 | 2741.00 |
| BL | 92016.00 | 80.14 | 150.39 | 0.00 | 3.00 | 19.00 | 91.00 | 5360.00 |
| HPALM | 60684.00 | 1.79 | 12.79 | 0.00 | 0.00 | 0.00 | 1.00 | 689.00 |
| HPALF | 60155.00 | 1.69 | 11.83 | 0.00 | 0.00 | 0.00 | 1.00 | 612.00 |
| HP | 65816.00 | 3.19 | 23.54 | 0.00 | 0.00 | 1.00 | 2.00 | 1301.00 |
| HIALM | 96036.00 | 74.63 | 124.08 | 0.00 | 7.00 | 27.00 | 92.00 | 2335.00 |
| HIALF | 95815.00 | 71.60 | 119.90 | 0.00 | 7.00 | 26.00 | 88.00 | 2214.00 |
| HI | 97024.00 | 144.58 | 242.30 | 0.00 | 13.00 | 52.00 | 178.00 | 4461.00 |
| TRALM | 91779.00 | 12.18 | 17.51 | 0.00 | 2.00 | 7.00 | 16.00 | 1790.00 |
| TRALF | 91594.00 | 11.83 | 17.76 | 0.00 | 2.00 | 7.00 | 16.00 | 1748.00 |
| TR | 94022.00 | 23.42 | 34.61 | 0.00 | 4.00 | 14.00 | 32.00 | 3538.00 |
| WHALM | 96368.00 | 120.11 | 141.32 | 0.00 | 21.00 | 86.00 | 169.00 | 6887.00 |
| WHALF | 96115.00 | 112.97 | 137.21 | 0.00 | 19.00 | 79.00 | 159.00 | 7410.00 |

**Data Preprocessing**

* Importing Data (Using Pandas)

Pandas is an open-source Python library that provides easy-to-use data structures and data analysis tools. It is widely used for data manipulation, analysis, and cleaning in the field of data science. Pandas allows users to work with two primary data structures: Series and DataFrame.

**Series:** A one-dimensional array-like data structure that can hold various types of data (e.g., numbers, strings, dates). Each element in the Series has an associated index, which can be customized or automatically generated.

**DataFrame:** A two-dimensional tabular data structure, similar to a spreadsheet or a SQL table, that consists of rows and columns. It allows users to store and manipulate data in a format that resembles a relational database table.

However, the data set that has been used here has more than 1million entries. To deal with such a big data Pandas is used.

* Dealing with Duplicates and Null Values

The dataset does not have any duplicate values. So, we move to the next step that is removing the Null values.

|  |  |
| --- | --- |
| X 0 | Y 0 |
| OBJECTID 0 | NCESSCH 0 |
| SURVYEAR 0 | STABR 0 |
| LEAID 0 | ST\_LEAID 0 |
| LEA\_NAME 0 | SCH\_NAME 0 |
| LSTREET1 3 | LSTREET2 100144 |
| LCITY 0 | LSTATE 0 |
| LZIP 0 | LZIP4 0 |
| PHONE 0 | CHARTER\_TEXT 0 |
| MAGNET\_TEXT 0 | VIRTUAL 0 |
| GSLO 0 | GSHI 0 |
| SCHOOL\_LEVEL 0 | TITLEI 0 |
| STITLEI 0 | STATUS 0 |
| SCHOOL\_TYPE\_TEXT 0 | SY\_STATUS\_TEXT 0 |
| ULOCALE 0 | NMCNTY 0 |
| TOTFRL 23758 | FRELCH 26605 |
| REDLCH 26605 | PK 69021 |
| KG 47185 | G01 46781 |
| G02 46742 | G03 46790 |
| G04 47000 | G05 48296 |
| G06 63502 | G07 68572 |
| G08 68344 | G09 73736 |
| G10 73875 | G11 73903 |
| G12 73991 | G13 100579 |
| UG 92569 | AE 100540 |
| TOTMENROL 100722 | TOTFENROL 100722 |
| TOTAL 2071 | MEMBER 2071 |
| FTE 9502 | STUTERATIO 1216 |
| AMALM 31433 | AMALF 31687 |
| AM 24373 | ASALM 17784 |
| ASALF 18097 | AS 13828 |
| BLALM 11599 | BLALF 12288 |
| BL 8706 | HPALM 40038 |
| HPALF 40567 | HP 34906 |
| HIALM 4686 | HIALF 4907 |
| HI 3698 | TRALM 8943 |
| TRALF 9128 | TR 6700 |
| WHALM 4354 | WHALF 4607 |
| WH 3784 | LATCOD 0 |
| LONCOD 0 |  |

From the table it can be seen that attributes ‘LSTREET2’, ‘TOTMENROL’, ‘G13’, ‘AE’, ‘TOTFENROL’ has more than 90% of its values as 0. So, these attributes can be deleted. Moreover, given attributes ‘TOTFRL', 'FRELCH' ,'REDLCH' ,'X', 'Y', 'OBJECTID', 'NCESSCH', 'PHONE' are of no use so these are deleted as well.

Grade values can be zero because number of students in different type of schools in particular class could be zero. For example, primary school won’t have any kindergarden student and vise versa. Therefore, It is acceptable if the NAN is replaced with 0. Moreover, there can be zero number of students from different race in some schools. so it can be Zero as well. So, wherever there is any NAN in given attributes 'KG' ,'G01', 'G02', 'G03', 'G04', 'G05', 'G06', 'G07', 'G08', 'G09', 'G10', 'G11', 'G12', 'G13', 'UG', 'AE', 'PK', 'AMALM', 'AMALF', 'AM', 'ASALM', 'ASALF', 'AS', 'BLALM', 'BLALF', 'BL', 'HPALM', 'HPALF', 'HP', 'HIALM', 'HIALF', 'HI', 'TRALM', 'TRALF', 'TR', 'WHALM', 'WHALF', 'WH' is replaced to 0.

The total number of students and members in one school cannot be zero. So, rows that have NAN in their total and Member Attribute are removed. So, the rows with zero in 'TOTAL', 'MEMBER', 'FTE', 'STUTERATIO' are deleted.

After dealing with all these Null values if we again check for the null values again we get:

|  |  |
| --- | --- |
| SURVYEAR 0 | STABR 0 |
| LEAID 0 | ST\_LEAID 0 |
| LEA\_NAME 0 | SCH\_NAME 0 |
| LSTREET1 0 | LCITY 0 |
| LSTATE 0 | LZIP 0 |
| LZIP4 0 | CHARTER\_TEXT 0 |
| MAGNET\_TEXT 0 | VIRTUAL 0 |
| GSLO 0 | GSHI 0 |
| SCHOOL\_LEVEL 0 | TITLEI 0 |
| STITLEI 0 | STATUS 0 |
| SCHOOL\_TYPE\_TEXT 0 | SY\_STATUS\_TEXT 0 |
| ULOCALE 0 | NMCNTY 0 |
| PK 0 | KG 0 |
| G01 0 | G02 0 |
| G03 0 | G04 0 |
| G05 0 | G06 0 |
| G07 0 | G08 0 |
| G09 0 | G10 0 |
| G11 0 | G12 0 |
| G13 0 | UG 0 |
| AE 0 | TOTAL 0 |
| MEMBER 0 | FTE 0 |
| STUTERATIO 0 | AMALM 0 |
| AMALF 0 | AM 0 |
| ASALM 0 | ASALF 0 |
| AS 0 | BLALM 0 |
| BLALF 0 | BL 0 |
| HPALM 0 | HPALF 0 |
| HP 0 | HIALM 0 |
| HIALF 0 | HI 0 |
| TRALM 0 | TRALF 0 |
| TR 0 | WHALM 0 |
| WHALF 0 | WH 0 |
| LATCOD 0 | LONCOD 0 |

**Discriptive Visuliztions**

The given Graph represents the number of schools present in different cities of Unitied States.

A graph of a number of school classes

Description automatically generated

It is noticeable that in year 2020 – 2021 California has the maximum number of schools with the count of 10000. Followed by Texas with 8453 number of schools.

A blue circle with green and orange triangles

Description automatically generated

The Pie chart here shows what percentage of the schools is charity and not charity. 88.0% of the schools are not charity and 7.3% are the charity one’s. However, 4.7% of the schools are unknown that they belong to which category.

A graph of different sizes and colors

Description automatically generated

Most of the schools took in person classes. There are some schools that offers Supplemental Virtual Classes. Which means the school take non virtual classes but there are some supplementary classes vich are provided virtually. The type Virtual with fase to fase options and Full Virtual schools are very low in count when compared to non virtual schools.

A graph of a number of students

Description automatically generated

There are almost 30,000 schools with Prekindergarten schools.

A graph of a number of students

Description automatically generated with medium confidence

Grade 5 classes are offered with the maximum number of 29987.

School types can be mentioned as for two categories.

1. For the level of school
2. For the type of education it is offering.
3. The schools with the different levels and their count over the united state can be seen from the given bar plot. From the plot we can see that the maximum number of schools that are there in united state are Elementary Schools, followed by High schools and then Middle schools. Moreover, there are other school types with a small number.

A graph of different types of school type

Description automatically generated

1. Most of the schools are regular schools. In comparison with the regular schools, alternative education schools, Special education Schools, Careers and technical schools are low in number.

A graph of different types of school type

Description automatically generated

**Correlation:**

A blue and red squares with white text

Description automatically generated

A correlation heatmap is a graphical representation used to visualize the correlation between different variables in a 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.

+3

**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.

**References**

[Public School Characteristics 2020-21 - Catalog (data.gov)](https://catalog.data.gov/dataset/public-school-characteristics-2020-21)

[Public School Characteristics 2020-21 - Original ISO-19139 metadata - Catalog](https://catalog.data.gov/dataset/public-school-characteristics-2020-21/resource/ec855cbe-51c0-4031-b09c-0fb6cf29e247)

[Public School Characteristics 2020-21 - ArcGIS Hub Dataset - Catalog](https://catalog.data.gov/dataset/public-school-characteristics-2020-21/resource/2e9a67a8-c231-4eb3-a800-0baff44b2761)

[Public School Characteristics 2020-21 - ArcGIS GeoService - Catalog (data.gov)](https://catalog.data.gov/dataset/public-school-characteristics-2020-21/resource/961912f5-514c-427b-add3-2a579e7caed3)

[Public School Characteristics 2020-21 - Public School Documentation - Catalog (data.gov)](https://catalog.data.gov/dataset/public-school-characteristics-2020-21/resource/612695c7-92e6-460f-8928-8fca83ad2157)

[Public School Characteristics 2020-21 - Public School District File - Catalog (data.gov)](https://catalog.data.gov/dataset/public-school-characteristics-2020-21/resource/d9edec94-f80f-440f-9916-b7a999555856)