**Project Background:** The provided dataset about Covid-19 describes the weekly time-series of positivity rates across the US universities spanning from September until November 2020. It includes external variables such as cities, county names, admission rates, number of students, total confirmed positive cases etc. The snapshot of the CSV file **COVID\_dataset\_v0.csv** with the dataset columns and rows are provided below:

|  |  |
| --- | --- |
| **A Data Science Competition**  **University of Illinois at Urbana-Champaign and California State University, Long Beach** | **Mining COVID-19 Testing Data: Understanding Behavior and Providing Actionable Insights**  – How can available data be used to develop a prediction model that accurately forecasts future positivity rates at the university-level? - What are the key factors identified through available data that can explain the variation in positivity rates at the university-level.  **Aigiun Guseinova, Almee Christian, & Taher Vora**  **August 11, 2023** |





Our Team **“Data Space”** aims at developing a prediction model that accurately forecasts future positivity rates at the university-level by identifying the key factors defined through the available data. Our findings should be able to explain the variation in positivity rates at the university level. All our analysis has been implemented in Google Collab, Jupiter Notebook and Tableau tools to gain insight into the research project to gain results.

**Google Collab URL:** [**COVID-19**](https://colab.research.google.com/drive/1HlvQVC_zqVZcRthgo4Lj5kUd2p-5JCoP#scrollTo=2dGAk0Vm9Mis)

The dataset is comprised of 1,130 observations and 18 characteristics with one dependent variable and 17 independent variables. We may encounter different types of data such as float, integer, string. No variable column has null/missing values.

**EDA (Exploratory Data Analysis)**: From the image below, it can be observed that there is a significant difference between the minimum value and the first quadrant (represented by 25%), and the 3rd quadrant (represented by 75%) and the maximum value, suggesting that there are extreme values (or \*\*outliers\*\*) in the dataset. We can also observe that mean > median for number\_students, TotalCount, TotalCountWeekly, TotaConfirmedPositive, TotalConfirmedPositiveWeekly, TotalPositivityRate and TotalPositivityRateWeekly, and the opposite for admission\_rate.

A colorful squares with black text

Description automatically generated with medium confidence

**Heatmap observations:** Here we can infer that TotalConfirmedPositive has a strong positive correlation with number\_students and a strong negative correlation with latitudewhere none of the fields correlate with latitude as much as they do with other fields.

**A computer screen shot of a computer error

Description automatically generated**

A group of blue and white bars

Description automatically generated

The largest indication of confirmed Covid-19 positive cases has been observed in the Los Angeles County in California being the largest number (32,563 cases), Dane County in Wisconsin (22,783 cases) and Brazos County (18,632 cases) in Texas.

A map of the united states with many colored circles

Description automatically generated

Our team has implemented three machine learning algorithms for regression predicting modelling: **Random Forest, XGBoost, and Decision Tree Regression models.**

In our case, **XGBoost Regression model** has an RMSE value of 3.039917 which indicates that, on average, the predictions made by your XGBoost regression model have an error of approximately 3.04 units that seems to be high for our model. This indication gives us a result our model is not accurate in predicting the target variable.

A screenshot of a computer

Description automatically generated

Our **Decision Tree regression model** has an average squared prediction error (MSE) of around 4.08. This suggests that, on average, the predictions of our model are reasonably close to the actual values, indicating a decent fit. An R² of approximately 0.79 indicates that around 79% of the variability in the target variable is explained by our Decision Tree regression model. A higher R² suggests that our model is capturing a significant portion of the variability in our data.

A screenshot of a computer

Description automatically generated

The provided values in the **Random Forest Regression** indicate that our regression model has a relatively low Mean Squared Error, which suggests a good fit of the model to the data with small prediction errors. R² value of approximately 0.975 suggests that about 97.6% of the variance in the dependent variable can be explained by the independent variables in our model. A high R² value indicates that our model is explaining a significant portion of the variability in the data with higher values indicating a better fit of the model to the data.

A screenshot of a computer

Description automatically generated

**Observations:**

- The below graph represents the positivity rate for different counties for the 3165th week (the last week of observations);

- It can be seen that the positivity rate has significantly increased and now the majority of the counties have positivity rate between 5 and 10.

**County names are as follows:**

Addison County, Alachua County, Alamance County, Alameda County, Albany County, Allegheny County, Androscoggin County, Baltimore city, Berkshire County, Boulder County, Brazos County, Broome County, Camden County, Centre County, Champaign County, Chittenden County, Clarke County, Cleveland County, Coles County, Cook County, Cuyahoga County, Dane County, Davidson County, DeKalb County, Delaware County, Denver County, District of Columbia, Durham County, Dutchess County, East Baton Rouge Parish, Erie County, Fayette County, Franklin County, Grafton County, Grand Forks County, Hampshire County, Hardin County, Harris County, Hartford County, Hennepin County, Jefferson County, Lancaster County, Leon County, Lexington city, Los Angeles County, Lyon County, Marion County, Marquette County, McLennan County, Mesa County, Miami-Dade County, Middlesex County, Milwaukee County, Mineral County, New York County, Norfolk city, Northampton County, Oakland County, Oktibbeha County, Onondaga County, Orange County, Orleans Parish, Payne County, Peoria County, Philadelphia County, Pickens County, Portage County, Providence County, Radford city, Rensselaer County, Richland County, Richmond city, Riley County, San Diego County, San Luis Obispo County, Santa Clara County, Sedgwick County, Spokane County, St. Joseph County, St. Louis city, Suffolk County, Tippecanoe County, Tolland County, Tompkins County, Travis County, Union County, Wake County, Washington County, Washtenaw County, Watauga County, Wayne County, Williamsburg city, Worcester County

