Danielle Soileau
Major in Biology
Mentor: Dr. Ashwini Kucknoor
Research in Microbiology/Molecular Biology
Department of Biology

Characterization of a novel cell surface protein coding gene, TfAD1 in *Tritrichomonas* foetus, a cattle pathogen.

The purpose of this project is to create regression models for the number of new daily cases and new daily deaths due to Covid-19 in each state in the US. These models can then be used to analyze the current pandemic as well as provide valuable information for future pandemics. The information provided by our models can allow people to make more informed decisions that has the potential to save more lives in the future. The models for this project were created and selected using a combination of statistical analysis and data science.

The data for this project was downloaded from John Hopkins University public dataset. The timeframe for the data is between Jan 23rd, 2020 and July 28th, 2021. We used different Python data modeling packages to model and analyze our data. The two major packages used for this project was Pandas and Sklearn. The main challenge of this project was designing a regression model that best fits the data. There are many regression models in literature. We chose linear, polynomial, decision tree, and support vector regression. We measured the efficiency of each model using the Mean Squared Error (MSE) and the Coefficient of Determination (R^2). Both the MSE and R^2 are a measurement of how close the model fits the data. The MSE is an absolute measurement and R^2 is a relative measurement. For MSE, the smallest value indicates the best fit. For R^2, the value closest to 1 indicates the best fit.

Results

The first model we developed is the linear regression model. A linear regression model is best used when the data seems to follow a straight line. Unfortunately, the COVID-19 data had many curves, therefore the linear model did not fit the data well. The linear model for the new daily cases for the state of Texas is shown in Figure 1.

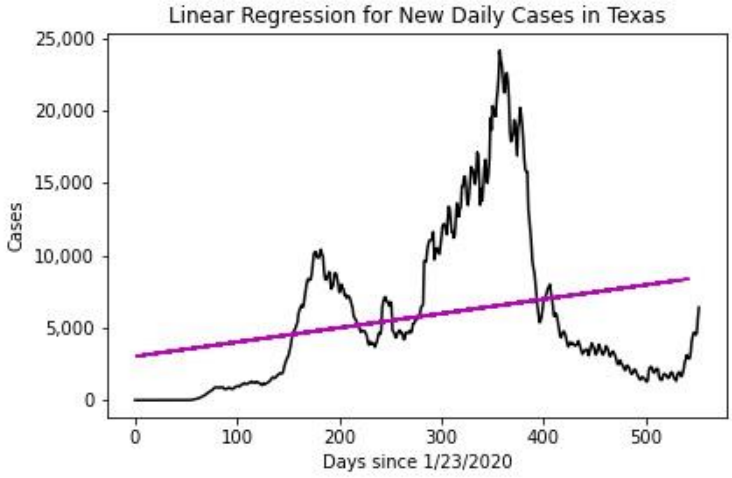

Figure 1: Linear Regression for Daily Cases in Texas

The second model developed is the polynomial regression model. We also introduced a rolling average to reduce the noise of our data. We took a rolling average of 14-days to create all subsequent models. The polynomial regression model was better than the linear model but still did not fit the data very well. The polynomial model for the new daily cases for the state of Texas is shown in Figure 2.

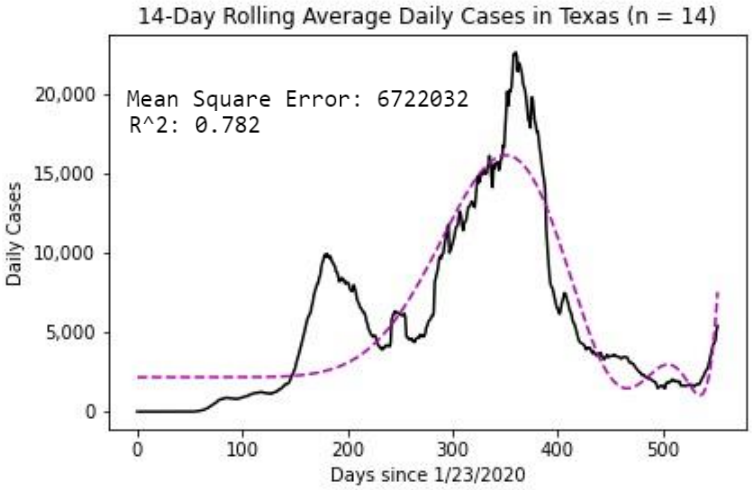

Figure 2: Polynomial Regression for Daily Cases in Texas

The next model was the decision tree regression model. This model turns the data into a tree with many different nodes. All nodes are broken down into two categories, decision nodes and leaf nodes. This model fitted our data better than linear regression and polynomial regression models, The decision tree model for the new daily cases for the state of Texas is shown in Figure 3. However, decision tree regression model might overfitting the data. One of the requirements for a good model is that it can predict other data sets in addition to the one that was used to create it. If a model is too reliant on one dataset it will not be useful for other datasets.

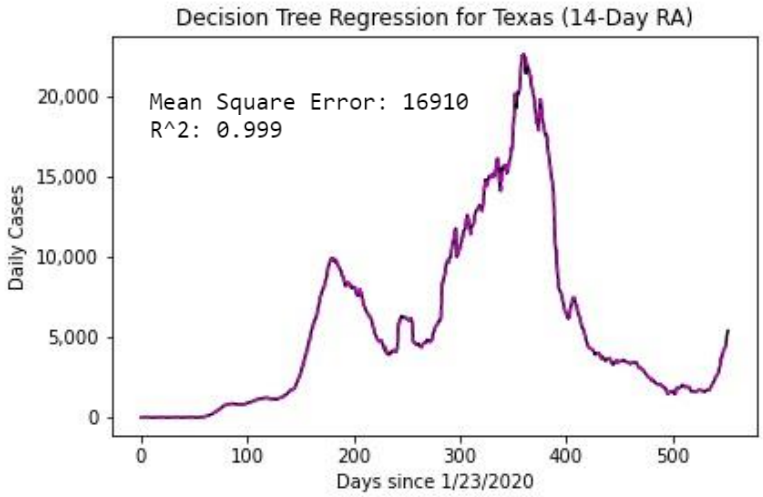

Figure 3: Decision Tree Regression for Daily Cases in Texas

The final model we developed was a support vector regression model. This model allows to ignore error within a certain threshold when creating a model. The points closest to the model are the support vectors and dictate what shape the model makes. The model tries to get as many data points as possible within that error threshold. This model was the best at predicting our data with the smallest MSE and R^2 closest to 1 compared with the other three models. The support vector models for the new daily cases and new daily deaths for the state of Texas are shown in Figures 4 and 5 respectively.

Conclusion

We were able to successfully develop four regression models for each chosen state in U.S.A. This research can be extended to include more countries besides the United States. It is also possible that there are even better regression models that fit the data better than the models used in this project. It would also be interesting to see if a different model would be better for the

data as the pandemic progresses. We plan to continue this project and present at the STEM and other academic conferences.

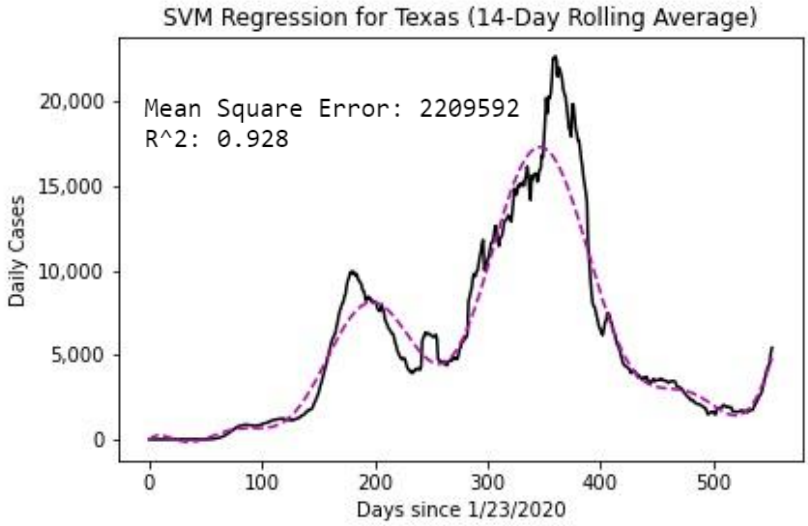

Figure 4: Support Vector Regression for Daily Cases in Texas

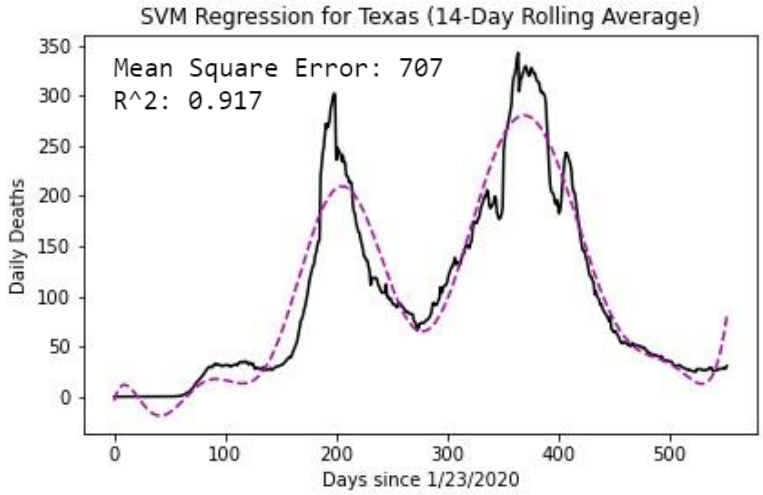

Figure 5: Support Vector Regression for Daily Deaths in Texas