Utilizing Machine Modeling to Predict Pathogenicity in Mutagens Leading to Hereditary Cataract

Elaine Liu[1], Noah Civiletti[2], Samuel Cook[3], Maria Kinzie[4],
Advisors: Dr. Tracy Chen[5], Dr. Yishi Wang[5], Dr. Ying Wang[5]
[1]Charlotte Latin School, [2]Northwood High School, [3]Highland School of Technology, [4]Davidson Early College High School, [5]University of North Carolina Wilmington

Introduction

Hereditary cataracts are a rare condition in which cataracts develop at an unusually young age as a result of mutations to the αB-crystallin protein. There is very little conclusive information surrounding what causes specific genetic mutations to be pathogenic. Thus, the goal of this project was to identify a machine learning model that could predict whether an αB-crystallin mutation was pathogenic through its biochemical characteristics and create a prediction tool for future researchers to utilize, while simultaneously gathering more information about the role of such characteristics in causing hereditary cataracts.

Variable Selection

In order to identify which variables would be the most useful at determining the pathogenicity of mutations, an exploratory data analysis was conducted. The EDA identified three main variables of interest: isoelectric point, solvent area, and evolution age as pdel. The isoelectric point is the pH value in which a protein has net electrical charge of zero. Next, relative solvent area measures the amino acid's accessibility to surrounding solvents (mostly water). Lastly, the evolutionary age as pdel is a standardized measurement of the amount of time an amino acid mutated from its original state.

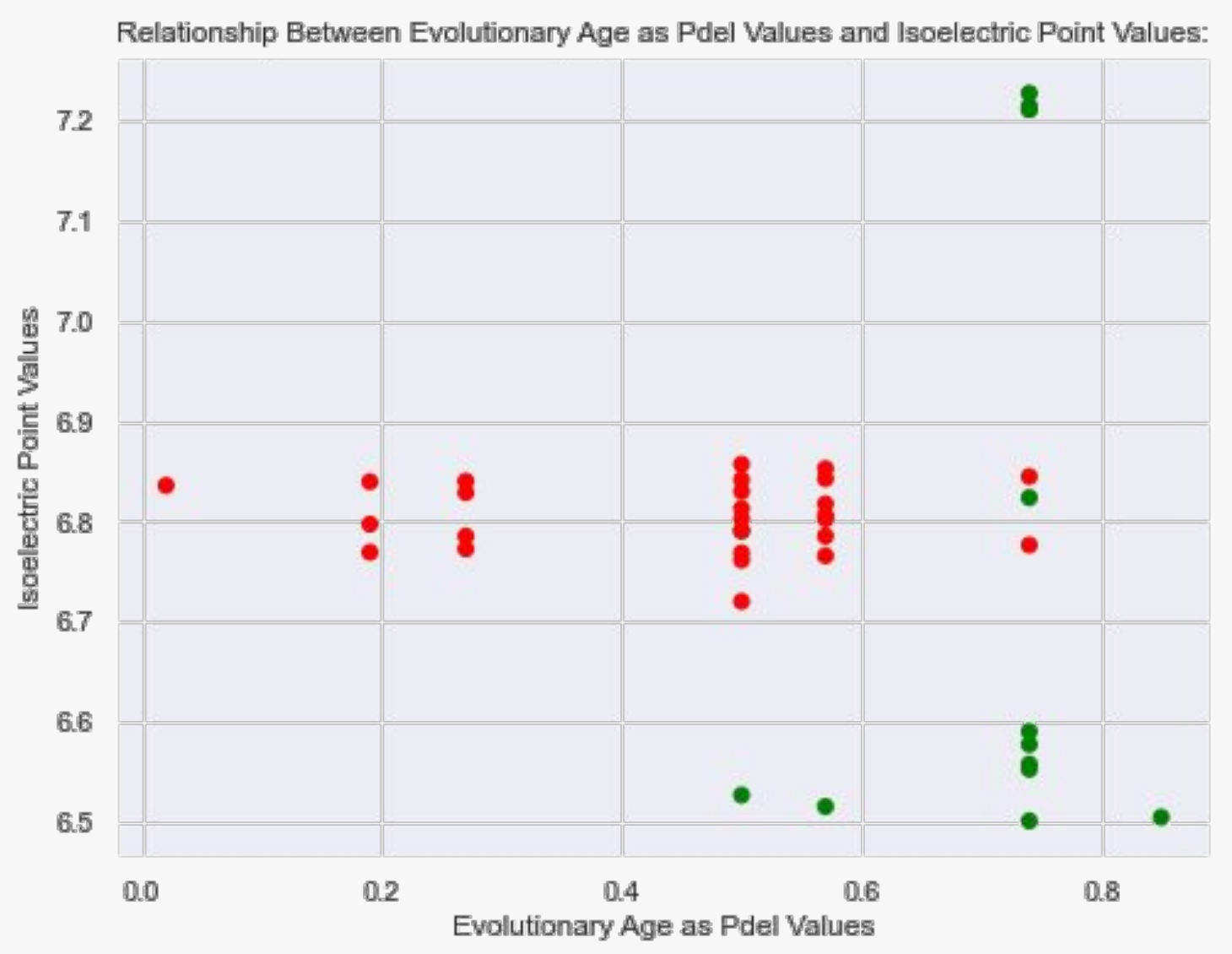

Figure 1

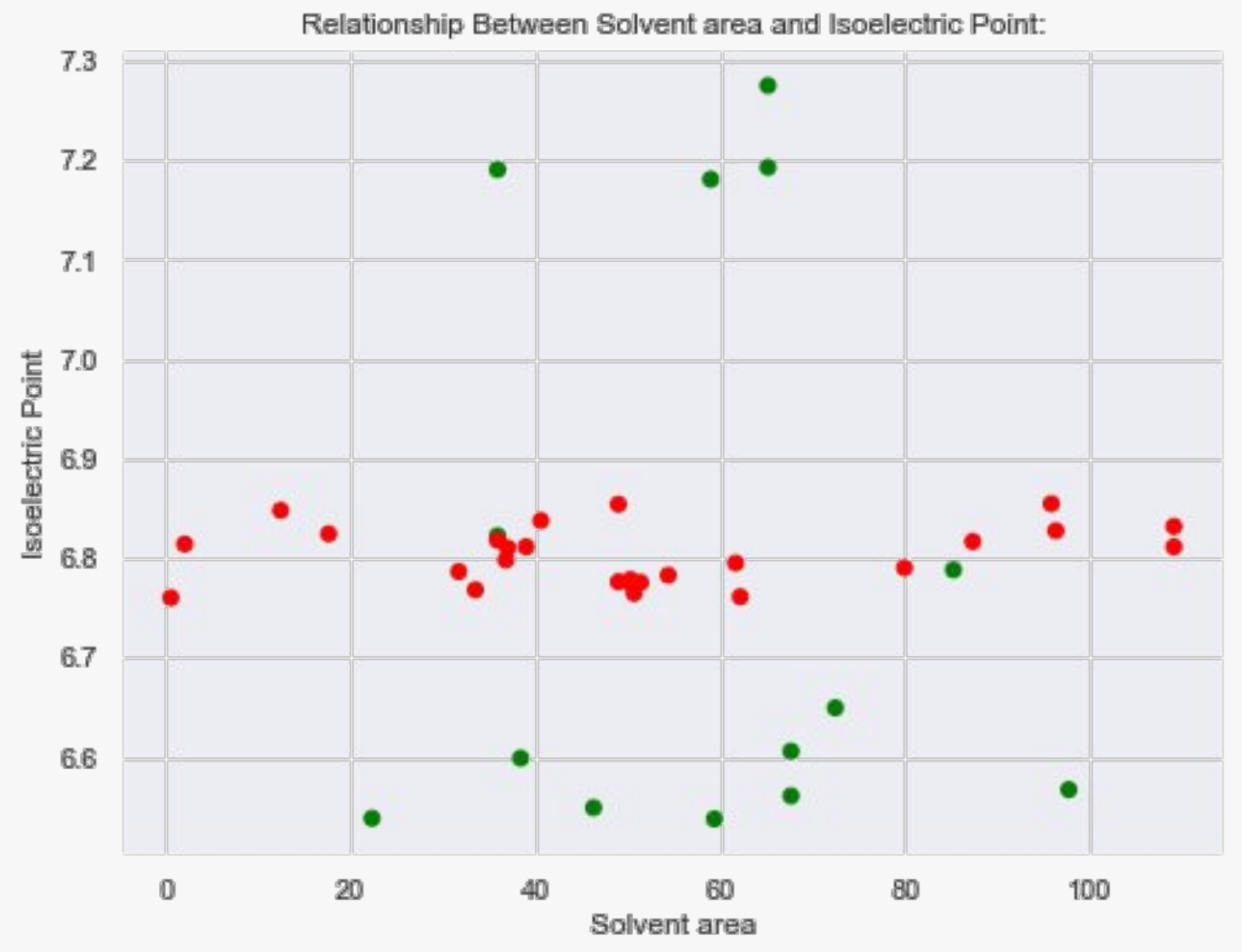

Figure 2

Negative values in red. Positive values in green.

Isoelectric point has the most predictive power out of all of the other variables. It clearly separates the data between the positive and negative groups in a nonlinear split. Besides a few outliers, the negative group's isoelectric points remain at about 6.76 (original state). Conversely, the positive group is separated into areas of higher and lower pH compared to the negative group; specifically the pH of the positive groups ranges from 7.2 to 6.5.

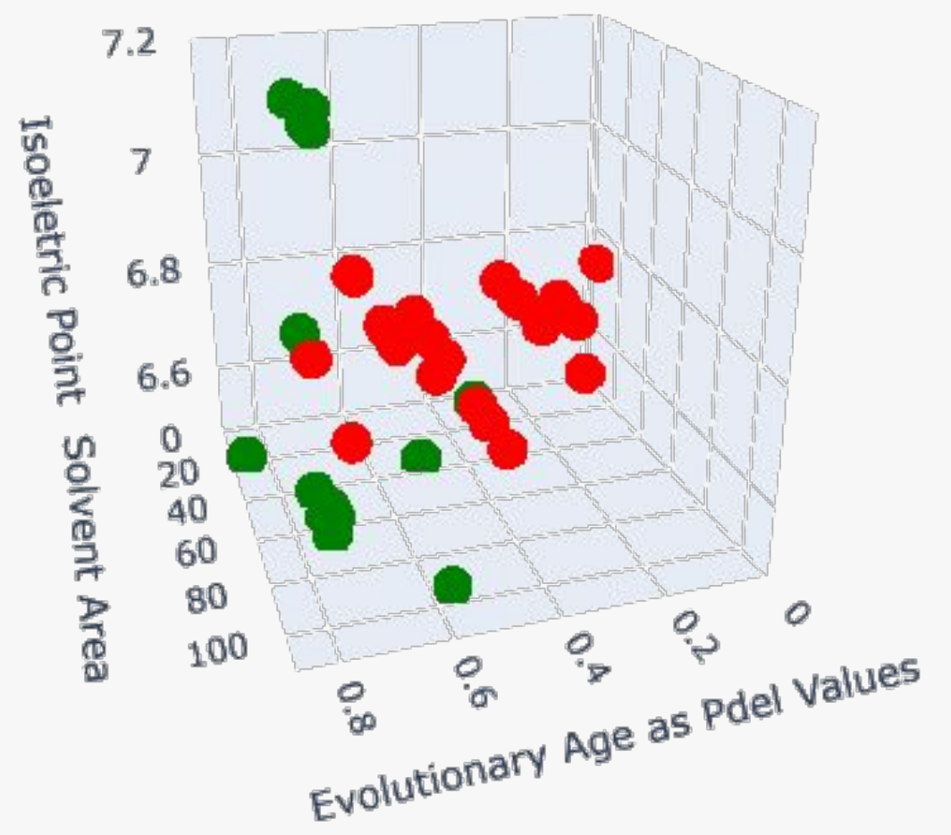

Figure 3

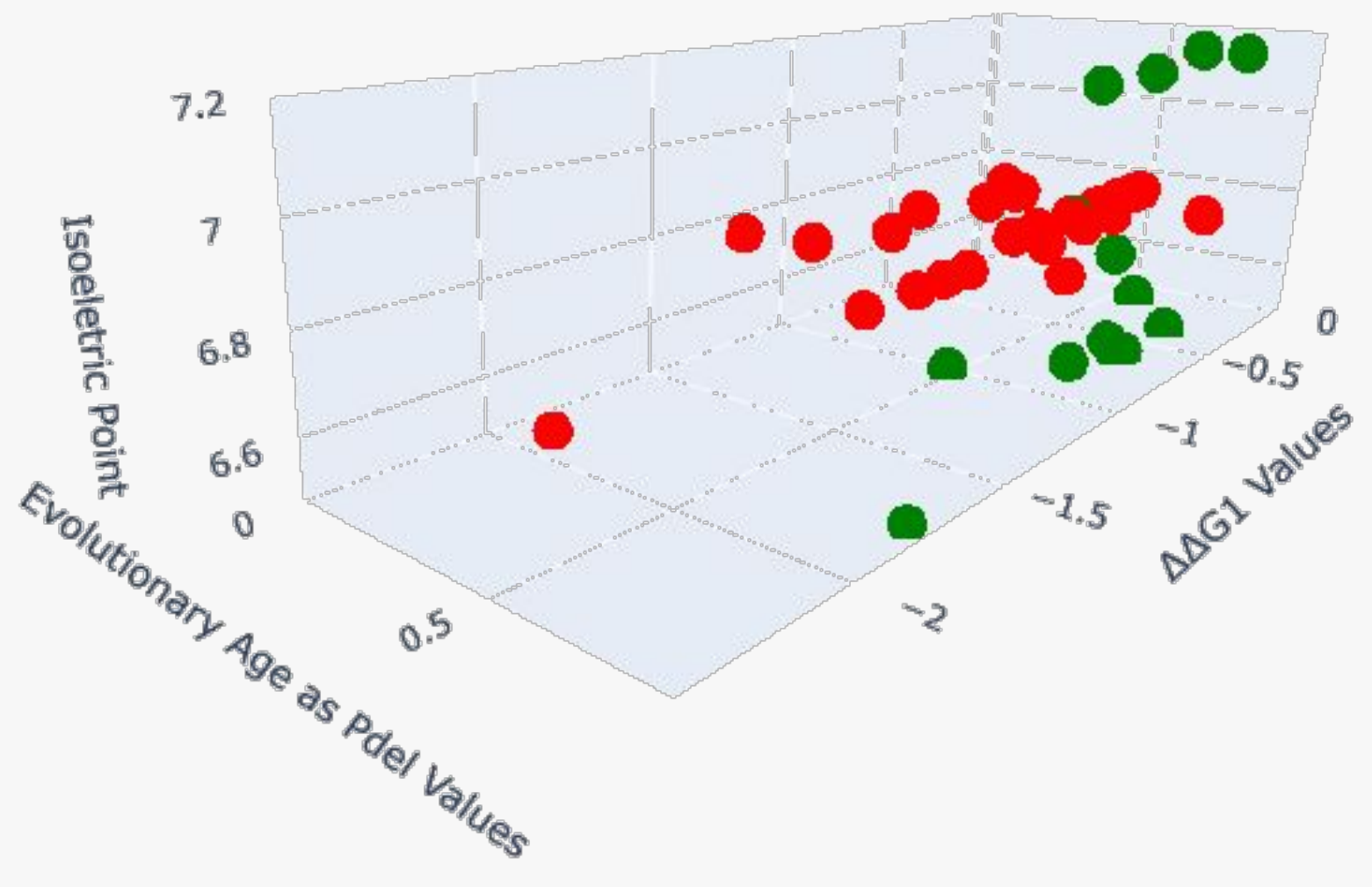

Figure 4

Negative values in red. Positive values in green.

Classifier Selection

In order to identify which classifier should be used in the final prediction model, a variety of different classifiers were tested against each other. The K-Nearest Neighbor, Adaptive Boosting, and Support Vector Classifier all had around equal accuracy. After further testing, SVC was selected due to its fast runtime and high predictability.

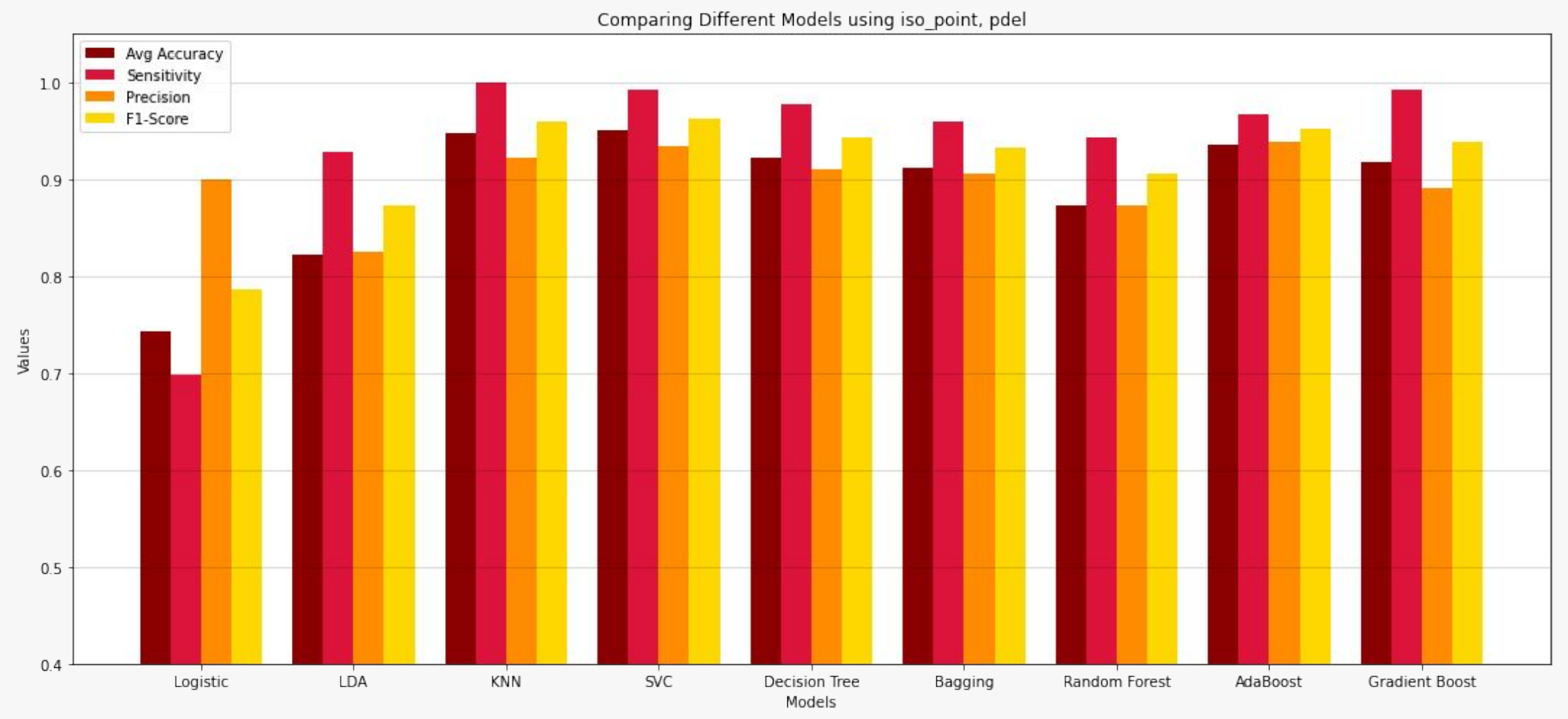

Figure 5

Support Vector Classifier

Support Vector Classifier models work by creating a hyperplane that is farthest away from data points on either side and separates the data into classes.

The kernel is a specific function used to find the dot product of the x and y vectors. Different data sets may produce more accurate results using different kernels that can better separate the data.

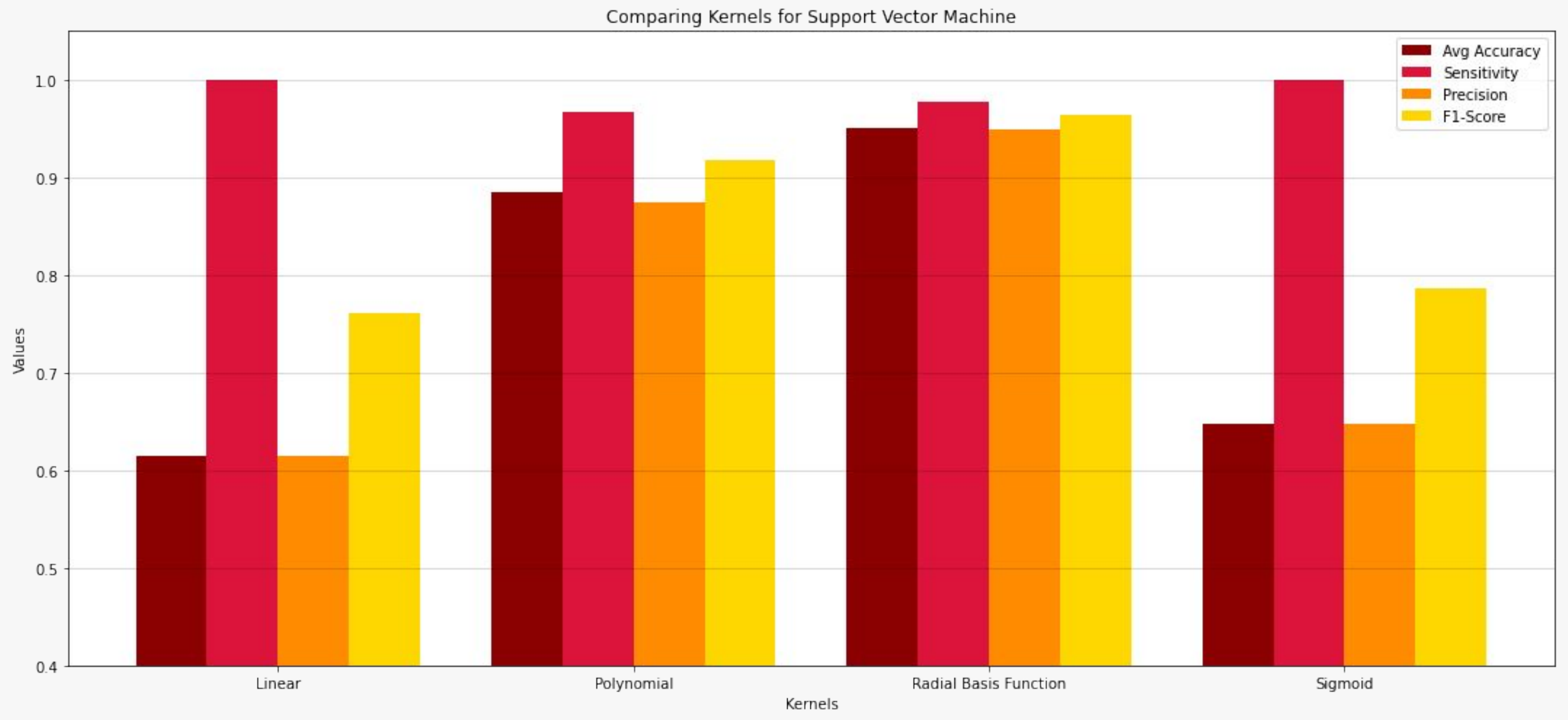

Figure 6

After experimentation the radial bias function (rbf) kernel was found to be optimal for both accuracy and F1-score. When applied to the correct variables, the SVC classifier had an average accuracy of around 96% with an F1-score of around 97%. Listed below are the top five performing SVC models.

SVC Models Tested with Different Covariate Combinations

Covariates	Avg Accuracy	SD Accuracy	Specificity	Precision	F1 Score	Run Time
['iso_point', 'solv_area']	0.965	0.053	1.000	0.951	0.975	0.095
['solv_area', 'iso_point']	0.953	0.063	1.000	0.932	0.965	0.087
['iso_point', 'pdel']	0.945	0.071	1.000	0.919	0.958	0.088
['iso_point', 'deldel_G2']	0.943	0.059	1.000	0.915	0.956	0.079
['pdel', 'iso_point']	0.940	0.062	0.992	0.920	0.955	0.076

Table 1

Results and Conclusion

Using the SVC classifier with isoelectric point and pdel as covariates, a prediction tool was created that would be able to predict the likelihood of a mutation being pathogenic given its biochemical characteristics. We tested our prediction tool with a new data set from the Biotechnology class. The results from this test are shown below:

Predictions for the pathogenicity of αB-crystallin mutations

Index	Prediction	Confidence Negative	Confidence Positive	True Value
R69C	positive	0.095	0.905	positive
D109A	positive	0.064	0.936	positive
P20S	negative	0.660	0.340	positive
R120G	positive	0.028	0.972	positive
A171T	negative	0.927	0.072	positive
R56W	positive	0.063	0.937	positive
D140N	positive	0.064	0.936	positive
D73N	positive	0.064	0.936	positive
R50L	positive	0.063	0.937	positive
R107L	positive	0.031	0.969	positive
R11H	positive	0.068	0.932	positive
R157H	positive	0.068	0.932	positive
I124V	negative	0.830	0.170	negative
S153T	negative	0.924	0.076	negative
L70M	negative	0.927	0.073	negative
S41A	negative	0.927	0.073	negative
S139T	negative	0.927	0.073	negative
T162S	negative	0.660	0.340	negative
E88D	negative	0.899	0.100	negative
I98V	negative	0.899	0.100	negative
S35T	negative	0.927	0.073	negative
V142I	negative	0.899	0.100	negative
L49F	negative	0.924	0.076	negative

Table 2

There were two false negatives for indexes P20S and A171T that were outliers of the isoelectric point. These outliers were consistent across every data set and were the reason why the machine learning models were not at 100% accuracy. However, overall, the model was still very accurate; pathogenicities were predicted with an average confidence of 0.93 for each observation.

The prediction tool shown above is in the public domain and is available for any researcher to use in the future.

Acknowledgements

This project was made possible by the guidance of Dr. Cuixian (Tracy) Chen, Dr. Yishi Wang, and Ms. Bailey Hall as well as Ms. Jessica Grey, Mr. Michael Suggs, Ms. Georgia Smith. Special thanks to Dr. Ying Wang, Ms. Jade Fluharty, and all of the students in the Summer Ventures at UNCW Biotechnology class for supplying the datasets and providing guidance.

References

Horwitz, J. (2003). Alpha-crystallin. Experimental Eye Research, 76(2), 145–153. https://doi.org/10.1016/s0014-4835(02)00278-6

de Jong, W. W., Leunissen, J. A., & Voorter, C. E. (1993). Evolution of the alpha-crystallin/small heat-shock protein family. Molecular biology and evolution, 10(1), 103–126. https://doi.org/10.1093/oxfordjournals.molbev.a039992