Determining the Impact of Socioeconomic Status on Athletic Performance Using Machine Learning

Zachary Hale* halezach11@gmail.com
Mikayla Kim* mikaylakim05@gmail.com
Andrew Qian* qiandrewj@gmail.com
Emma Savov* emmassavov@gmail.com
Ria Shah* rias0906@gmail.com
Thea Spellmeyer* tspellmeyer05@gmail.com
Julia Ma† julia.e.ma@lmco.com

*Governor's School of New Jersey Program in Engineering & Technology
†Corresponding Author
Zachary Hale, Mikayla Kim, Andrew Qian, Emma Savov, Ria Shah, & Thea Spellmeyer all contributed to this project equally.

***Abstract*—Because increasing athletic competition at the middle and high school levels has shifted the focus in sports towards achieving superior physical performance, optimizing athleticism and minimizing injury risk have become even more critical. However, the effects of socioeconomic disparity extend into youth sports participation; using machine learning algorithms, this research examines the correlations between socioeconomic statistics among school districts and the physical ability scores of youth athletes in those communities. Although all of the models achieved predictive accuracies above 94.2 percent when determining athleticism scores, they ultimately indicate that there is no direct correlation between a school district's socioeconomic profile and the physical ability of its students.**

I. Introduction

Due to increasing competitiveness in youth sports, a high level athlete's value to sports teams now depends on their physical performance and injury risk profile. Sparta Science is a sports technology company that has developed an algorithm to evaluate such factors in an individual. Leveraging measurements from force plate scans, Sparta Science assigns an individual a Sparta Score that indicates their level of fitness and musculoskeletal efficiency [1].

Developments in the sports industry continue to further athletic advancement. Yet, class disparities remain prevalent in youth athletics. In the United States, a lower socioeconomic status and identification with an ethnic minority aligns with a lower rate of involvement in sports: children with household incomes above $100,000 see sports participation rates twice those of children with household incomes below $25,000 [2]. Although there is an upward trend in youth sports participation among wealthy families, participation has decreased among lower income families from 42 percent in 2011 to 34 percent in 2017 [3].

The goal of this research is to investigate potential correlations between socioeconomic status and athletic performance. Athleticism is not only influenced by natural ability but also overall physical fitness, healthcare coverage, and access to advanced training opportunities. By applying machine learning algorithms to a data set containing mean Sparta Score values for school districts, as well as statistics regarding the districts' socioeconomic status, models were created to predict Sparta Scores. This paper discusses the development of these models and the real-world implications they have to highlight the effects that class has on athletic performance.

II. Background

A. Sparta Score

The Sparta Score is a metric determined by Sparta Science's proprietary algorithm that provides a holistic evaluation of an individual's movement and their levels of force production based on body scans. With comprehensive force plate data, Sparta Science calculates Load, Explode, and Drive subscores, representing different types of jumping movement, which are then aggregated to form the Sparta Score (see Figure 1) [4]. An athlete's subscores also indicate their injury risk due to musculoskeletal imbalances, taking into account their genetics, movement signature, and training history; a higher Sparta Score is more desirable and represents superior movement efficiency. For both members of the National Football League (NFL) Alumni and participants in its football camps, the NFL Alumni conducts Sparta Science scans on participants over time to track their physical progress.

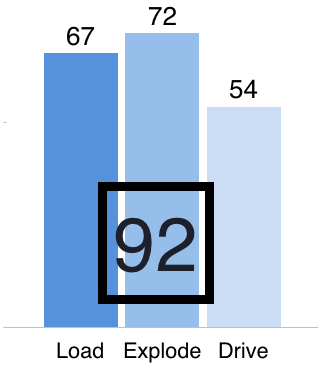

Fig. 1. Composite Sparta Score of 92 assigned based on subscores [5]

978-1-6654-7345-3/22/$31.00 © 2022 IEEE

B. Machine Learning

Machine learning is a branch of artificial intelligence that enables computers to optimize data analysis algorithms independently. It emulates how humans learn by aggregating past data, and is more efficient than humans at analyzing relationships between many variables. Machine learning models enable both classification and regression of sample data. Therefore, it is integral to data analysis to demonstrate trends between parameters and an output value. Supervised learning, a subset of machine learning, is an approach that trains algorithms on labeled data sets to make predictions. Thus, a supervised learning approach to data analysis allows for accurate, defensible outputs by analyzing data with existing patterns and trends.

1) Random Forest Regressors: In a decision tree algorithm, a branching series of binary statements uses various data parameters to predict the value of a target variable. The random forest algorithm is an amplification of the decision tree: predictions made by multiple decision trees are combined to output a final result. This integration of various predictive models, known as ensemble learning, results in a model with increased accuracy. Furthermore, random forest models reduce the risk of overfitting, where a machine learning algorithm fits too closely to the training data set and thus no longer makes effective predictions for testing sets or the entire data set.

2) Gradient Boosting: Gradient boosting is a machine learning model that serves as a valuable optimization method when implemented in conjunction with other predictive algorithms. Gradient boosting algorithms aggregate successive decision trees to minimize predictive error with a calculus procedure called gradient descent, which converges on the local minimum of a function [6].

C. Model Optimization

1) Hyperparameters: Hyperparameters control the learning process of a machine learning algorithm. Tuning hyperparameters is a critical step to optimizing the accuracy of a predictive model and preventing overfitting. By iteratively adjusting the hyperparameters of a machine learning algorithm, a model with higher accuracy rates is produced. However, a balance must be struck between the continuous optimization of hyperparameters and the requisite time and processing loads.

2) Statistical Analysis: In supervised learning applications, mean absolute error (MAE) and mean absolute percentage error (MAPE) indicate the accuracy of a machine learning model's predicted value set in comparison to the actual value set found in the data set. Moreover, these metrics warn against overfitting: a large deficit in the accuracy of the algorithm as applied to the testing set versus the training set demonstrates that the model is likely fitted too closely to the training set [7].

III. Procedure

A. Data Acquisition

The NFL Alumni provided two data sets with Sparta Science test information for participants in their youth fitness programs. The first contained 11,808 data points, with empirical force pad data for 121 subjects engaging in jump, balance, plank, and landing tests. The second contained 10,062 data points across 781 subjects, who received calculated scores for their performance on the tests. Both data sets assigned a holistic Sparta Score to subjects participating in the jump test.

Further research was conducted to match each participant in the data sets with their school district, verifying information found on sports profile websites with age and weight data. For each school district, the following statistics were collected from the National Center for Education Statistics's 2015 to 2019 database: the proportion of White residents; the percentage of families who only speak English at home; the percentage of residents with a disability; the percentage of residents with health insurance coverage; the percentage of families with incomes below the poverty line; the percentage of families accessing Supplemental Nutrition Assistance Program benefits; the percentage of married-couple households; the median household income of parents of children in public school; the unemployment rate; and the educational attainment rates of parents in the district [8]. These categories were chosen to thoroughly portray a school district's socioeconomic condition and explore potentially novel causal relationships at the intersection of socioeconomic status and physical fitness.

B. Data Preprocessing

1) Data Cleaning: The data sets provided by the NFL Alumni were first combined and cleaned to remove subjects without a Sparta Score, subjects lacking school district information, and duplicate subjects. In order to mitigate overrepresentation by a select few districts in the data set, districts with more than one participant received a mean Sparta Score to represent the average performance of athletes in that community. At the end of the data cleaning process, the data set contained 924 data points, with 78 unique school districts and 11 statistics per sample.

2) Manual Data Analysis: Before feeding the cleaned data set through the machine learning algorithms, it was manually analyzed in Google Sheets. Ten scatter plots comparing each average school district statistic to participants' Sparta Scores were created to discover preliminary correlations or the lack thereof. From these scatter plots, the slopes of the lines of best fit and the correlation coefficients were analyzed for direct or inverse proportionality between socioeconomic status and athletic performance.

C. Model Construction

1) Random Forest Algorithm: An 80-20 train-test split was created in the data set. The training set was passed into a random forest algorithm containing 1,000 unique decision trees in order to develop a predictive model for Sparta Scores. The same model was then applied to the testing set, returning a list of predicted Sparta Scores.

With the baseline random forest model established, further analysis was completed in order to determine which

parameters contributed most significantly to the Sparta Score prediction. A method was written to examine each individual decision tree and return a list of variables with above-average prevalence in the random forest. Then, the random forest regressor was executed using only the selected variables. Gradient boosting was employed as a third model for predicting the average Sparta Score of a school district based on its socioeconomic statistics.

D. Model Optimization

1) Hyperparameter Tuning: In order to further increase the accuracy of the random forest models, hyperparameter tuning was implemented. A grid composed of each of the various hyperparameters applicable to the random forest regressor was created and used with the GridSearchCV function from scikit-learn to return the most accurate model, which systematically tests every combination of hyperparameters in a set.

The array of hyperparameters tested included the following: n_estimators, the number of decision trees in the random forest, from 200 to 2,000 in intervals of 10; max_features, the number of variables to consider, chosen between the total number of parameters or its square root; max_depth, the maximum number of levels in the tree, ranging from 10 to 110 in intervals of 11; min_samples_split, the minimum number of samples necessary to make a decision, chosen between 2, 5, and 10; min_samples_leaf, the minimum number of samples required at an end node, chosen between 1, 2, and 4; and bootstrap, a true or false value which decides the method of selecting samples for model training.

This set of hyperparameters was further expanded by the number of folds tested, which splits the data into subsets. The GridSearchCV procedure iterated through each combination of hyperparameters; the grid contains 4,320 unique combinations for the random forest models, and each combination was tested on 2 folds. Since bootstrap was not a modifiable hyperparameter for the gradient boosting model, 2,160 existing possible combinations were tested with 2 folds each.

IV. RESULTS

The data set, with 78 samples and a confidence level of 95 percent, has a confidence interval between 75.308 and 78.259. This represents the range of values that contains the mean of the Sparta Scores in the data set 95 percent of the time. The preliminary data analysis in Google Sheets demonstrates a lack of clear correlations between athletic performance and socioeconomic status because the r^2 values and slopes of the lines of best fit for each scatter plot were close to zero (see Table I). Even before machine learning models were implemented, these graphs suggested that there were no clear correlations between socioeconomic statistics and Sparta Scores. The implementation of machine learning to analyze the data set further supports this hypothesis that little to no correlation exists.

TABLE I
TREND DATA FOR GRAPHS PLOTTING AVERAGE SPARTA SCORE AGAINST VARIOUS SOCIOECONOMIC STATISTICS

Independent Variable	r^2	**Best Fit Slope**
% White Residents	0.005	0.0186
% Speaks English Only	0.006	0.0293
% Disability	0	0.0249
% Health Insurance Coverage	0.004	-0.1300
% Below Poverty Line	0.002	-0.0008
% Food Stamp Benefits	0.002	0.0249
% Married-couple families	0	-0.0040
Median Household Income	0.01	0.0251
% Unemployment	0.01	0.0251
% High School	0.014	0.0426

All three machine learning models are highly accurate at predicting the average Sparta Score of a school district based on its socioeconomic statistics. As shown in Tables II, III, and IV, hyperparameter tuning further minimized predictive error for all three machine learning models on the testing set. This high accuracy corroborates the plausibility of conclusions drawn from the decision tree algorithms because the models closely represent the true conditions of the data set.

TABLE II
ERROR TABLE: RANDOM FOREST REGRESSOR WITH ALL PARAMETERS

	Before Tuning	**After Tuning**
Training MAE	1.7	2.7
Training MAPE	2.3%	3.8%
Testing MAE	4.7	4.6
Testing MAPE	5.9%	5.8%

TABLE III
ERROR TABLE: RANDOM FOREST REGRESSOR WITH SELECT PARAMETERS

	Before Tuning	**After Tuning**
Training MAE	1.8	3.4
Training MAPE	2.5%	4.8%
Testing MAE	5.2	4.0
Testing MAPE	6.6%	5.1%

TABLE IV
ERROR TABLE: GRADIENT BOOSTING MODEL

	Before Tuning	**After Tuning**
Training MAE	0.0064	0.59
Training MAPE	0.087%	0.83%
Testing MAE	5.1	4.5
Testing MAPE	6.5%	5.7%

Analysis of individual decision trees in the first random forest model indicates that while the machine learning algorithm achieved high predictive accuracy, the decisions it took to develop the regression remained arbitrary: there are no straightforward correlations between socioeconomic parameters and a school district's average Sparta Score. Figure 2 demonstrates that while the percentage of residents with disabilities has a positive proportionality with Sparta Score in one branch, it has the opposite relationship in the next branch.

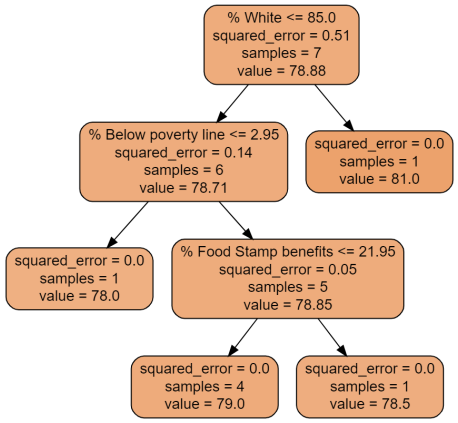

Fig. 2. Leaf nodes snipped from the decision tree with all 10 socioeconomic variables

These contradictory patterns support the finding that there is no direct correlation between any individual socioeconomic statistic and athletic performance. Analysis of the manually created scatter plots, which fail to indicate strong variable relationships, strengthens the conclusion of a negligible trend (see Table I).

The second machine learning model was a similar random forest regressor, but only with the following input parameters: the percentage of residents with a disability, the unemployment rate, and the percent of adult residents who had only attained education at the high school level or lower. These variables were chosen for the algorithm because they were most prevalent in the previous model.

Despite narrowing the scope of input variables in the random forest algorithm from 10 to 3, the second model retains a similar level of complexity with respect to the number of branching layers observed in the decision tree. Moreover, despite the presence of only 3 parameters in this random forest model, there remains a lack of clear correlations associating socioeconomic statistics with athletic performance. Figure 3 shows another instance of contradictory relationships between a district's disability percentage and its average Sparta Score.

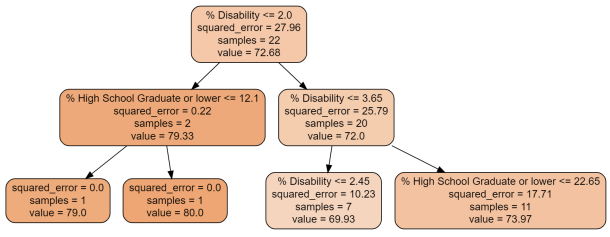

Fig. 3. Leaf nodes selected from the decision tree using only three socioeconomic variables

Of the three metrics—percentage of residents with a disability, unemployment rate, and the percent of adult residents only attaining a high school diploma or lower—none consistently demonstrated positive or negative proportionality with athletic performance. Thus, the most probable conclusion is that there is no correlation between socioeconomic status and athleticism.

When iterating through the gradient boosted model, four variables were represented most prevalently: the percentage of residents who had only graduated from high school or lower, the percentage of families with health insurance coverage, the proportion of residents with disabilities, and the unemployment rate. These statistics demonstrate very low correlation coefficients and trend line slopes when plotted against a district's average Sparta Score with a linear regression (see Table I). Therefore, these variables have a very weak relationship with athletic performance. The analysis of the manually created scatter plots and machine learning algorithms both support this conclusion.

V. DISCUSSION

Analysis of both the models and data suggests that there is no correlation between the socioeconomic status and an athlete's Sparta Score. However, external research contradicts this conclusion, suggesting that a high socioeconomic background relates to high athletic performance. An explanation for the lack of expected trends is that the data set does not accurately represent the full range of socioeconomic backgrounds in the United States: the $125 cost of the Sparta Scan excludes low affluence individuals from the set.

Increased access to health care contributes to the positive correlation between socioeconomic status and athleticism. Receiving higher quality medical care promotes a faster and more complete recovery from injuries, allowing athletes to advance [9]. Conversely, athletes who are unable to comfortably afford medical attention are more likely to postpone or decline recommended treatments such as surgeries or MRI scans, leading to deteriorating physical performance [10].

The increased access to resources and opportunities is reflected in areas beyond health care: affluent school districts invest more in their sports programs, improving the quality and quantity of opportunities for young athletes. The ability to exercise with high-quality gym equipment promotes muscle growth and conditions the body to reduce injury risk [11]. Research also indicates that greater investment in athletic programs leads to a higher sports participation rate [12]. Therefore, higher income athletes will score higher on athletic performance tests such as the Sparta Scan because they receive better opportunities to pursue athletic development.

Low-income students face many limiting factors that interfere with their participation in athletic activities, including the lack of transportation, high equipment costs, and the inability to afford the risk of injuries [13]. Children may also be encouraged to find employment and help support their families instead of participating in sports. These factors negatively affect an individual's athletic development.

VI. CONCLUSIONS

A. Synopsis

The data analysis demonstrates a lack of a direct correlation between a school district's socioeconomic profile and its ath-

letic ability. This contradicts existing research that associates higher socioeconomic status with increased participation and improved performance in sports. However, this finding is promising since it encourages athletic participation among adolescents from all socioeconomic backgrounds. Even if there is little effect of an individual's social status on their potential for athletic excellence, advanced training programs and bodily health scans should be made readily accessible to promote higher standards of competition and the economic growth of the sports industry.

B. Limitations

As discussed earlier, the Sparta Scores for individual participants from certain school districts were condensed into one mean Sparta Score per district to resolve the issue of overrepresentation. Although this step succeeded in resolving sampling bias and created a more accurate profile of various school districts in comparison to each other, it significantly reduced the number of samples in the dataset. Thus, there were fewer data points to train the models and illustrate trends.

Further limitations to this dataset arose during the data preprocessing phase. Individuals without readily available school district information from online sports profiles were excluded from the machine learning models and subsequent analysis. Therefore, systemic disparities, such as Internet access, that might prevent an athlete from being discoverable online have an outsized influence on the examination of correlations between socioeconomic status and athletic performance.

Lastly, assumptions made in the collection of socioeconomic data limit the scope of analysis that may occur with this dataset. Because only average statistics were collected per school district, this precludes the possibility of more individualized investigation that may establish trends between class and athletic performance at the household level.

C. Future Work

1) Accessibility: The high cost of Sparta Scans is prohibitive to underprivileged families. By making scans more accessible, the machine learning models could more accurately reflect the diverse socioeconomic situations within the American population.

2) Private Schools: Private schools were unable to be factored into this research because the demographic makeup of their students is not publicly available. Private schools range widely in opportunity and resources compared to public schools, so including them will form a more comprehensive assessment of the correlations between socioeconomic data and athletic performance.

4) Individual Income Data: Median income is an important variable when investigating correlations involving socioeconomic status. Due to the potential error in generalizing a players income based on their school district, recording a more accurate income for each player would greatly improve the models.

3) Detailed Healthcare: While access to healthcare was a variable in the models, more detailed data reflecting access to specialty physicians and insurance quality will show the ability of someone to access comprehensive healthcare resources. This will aid our models in detecting patterns between healthcare and athletic performance.

ACKNOWLEDGMENTS

The authors of this project gratefully acknowledge the following for their assistance and guidance in the completion of this project: Rutgers School of Engineering, Rutgers University, and the New Jersey Office of the Secretary of Higher Education; Lockheed Martin and the NFL Alumni; Governor's School alumni; Dean Jean Patrick Antoine; Project Mentor Julia Ma; Residential Teaching Assistant Sakshi Lende; NFL Alumni Performance Lab Director Dr. Chuck Morris; Lockheed Martin Project Volunteers Marissa Delrocini, Nicholas Maranca, Lezly Murphy, and Brittany Yesner; Head Residential Teaching Assistant Ian Joshua Origenes; and lastly, Research Coordinator June Lee.

REFERENCES

[1] "Sparta Scan Overview: Jump. Balance. Plank." Sparta Science.

[2] A. W. Kuhn, A. Z. Grusky, C. R. Cash, A. L. Churchwell, and A. B. Diamond, "Disparities and Inequities in Youth Sports," Current Sports Medicine Reports, vol. 20, no. 9, pp. 494–498, Sep. 2021, doi: 10.1249/JSR.0000000000000881.

[3] D. Thompson, "Income Inequality Explains the Decline of Youth Sports," The Atlantic, Nov. 06, 2018. https://www.theatlantic.com/ideas/archive/2018/11/income-inequality-explains-decline-youth-sports/574975/

[4] "The Sparta Score - How Do You Stack Up?," Sparta Science, May 06, 2019. https://spartascience.com/the-sparta-score-how-do-you-stack-up/ (accessed Jul. 16, 2022).

[5] Jason Brownlee, "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning," Machine Learning Mastery, Nov. 20, 2018. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

[6] "ACS School District Profile 2015-19," nces.ed.gov. https://nces.ed.gov/Programs/Edge/ACSDashboard

[7] A. L. Duca, "How to check if a classification model is overfitted using scikit-learn," Medium, 28-Sep-2021. [Online]. Available: https://towardsdatascience.com/how-to-check-if-a-classification-model-is-overfitted-using-scikit-learn-148b6b19af8b. [Accessed: 05-Sep-2022].

[8] S. H. Woolf, L. Y. Aron, L. Dubay, S. M. Simon, E. Zimmerman, and K. Luk, "How Are Income and Wealth Linked to Health and Longevity?," Urban Institute, Jun. 04, 2016. https://www.urban.org/research/publication/how-are-income-and-wealth-linked-health-and-longevity

[9] Progressive Physical Therapy, "Articles," progressive-pt. https://www.progressive-pt.net/general/i6nyvqwo13/Early-Physical-Therapy-Can-Decrease-Recovery-Time

[10] S. Allin, C. Masseria, and E. Mossialos, "Measuring Socioeconomic Differences in Use of Health Care Services by Wealth Versus by Income," American Journal of Public Health, vol. 99, no. 10, pp. 1849–1855, Oct. 2009, doi: 10.2105/ajph.2008.141499.

[11] "Benefits of Gym Equipment for Physical Fitness," Physique Sports, Oct. 26, 2015. https://www.physiquesports.co.uk/blog/benefits-of-gym-equipment-for-physical-fitness/

[12] J. Grabmeier, "Want to play college sports? A wealthy family helps," Ohio State News, Aug. 30, 2021. https://news.osu.edu/want-to-play-college-sports-a-wealthy-family-helps/

[13] P. S. Tandon, E. Kroshus, K. Olsen, K. Garrett, P. Qu, and J. McCleery, "Socioeconomic Inequities in Youth Participation in Physical Activity and Sports," International Journal of Environmental Research and Public Health, vol. 18, no. 13, p. 6946, Jun. 2021, doi: 10.3390/ijerph18136946.