Daniel Stanley

Height Prediction Paper

Abstract:

With the use of an accurate height prediction, doctors understand a patient better by seeing if they are under or above their predicted height. I used basic data sets to see if some learning models have a chance of predicting height based on country and age. Decision tree regressor, Linear regression, and random forest regressions were successful models that I used. The raw AI prediction varied between a miss of 40cm all the way to 100 cm throughout all models. After hyper tunneling the model, it got to a closely stable average of 17 cm off the real height. Although being off 17 cm is not respectable in the Medical World, this proves that with higher ends of data, height prediction may be applicable to medical use.

Introduction

My research revolved around understanding if AI can be applied to predicting height. The prediction of height can lead to a number of different breakthroughs such as a better understanding of growth and hormone deficiency. With a better understanding of a child's growth rate, doctors can correctly diagnose patients earlier on in their life before it becomes too late. This type of problem is a supervised problem. Mainly because my model is going to be predicting height this problem would fit into the regression category. My data was numerical data, which originally included mean height, age, and year. But after further understanding, the year was removed. The end goal being to understand if A.I can be used in this area of field, if so what basic models work the best.

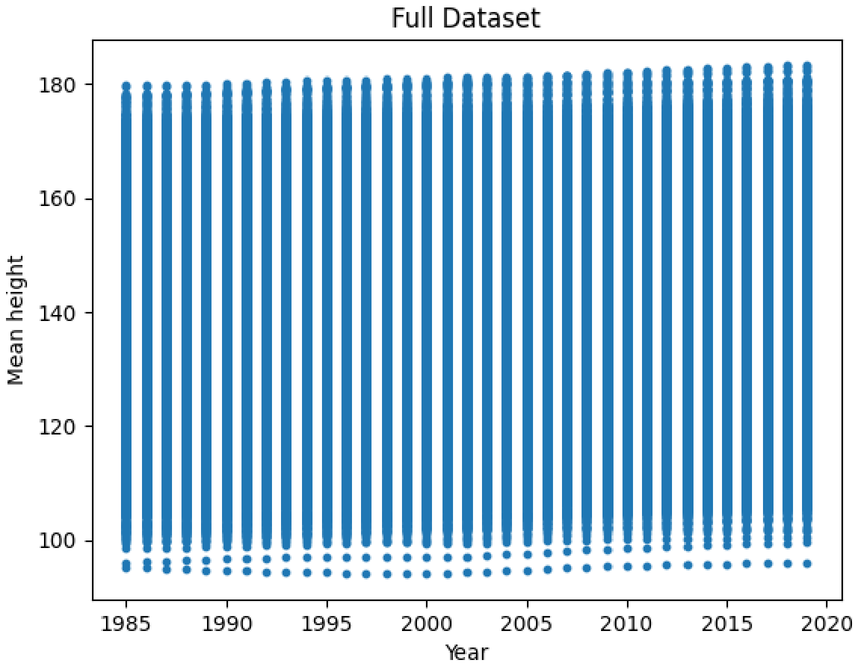
Background

DETERMINANTS OF VARIATION IN ADULT BODY HEIGHT by KARRI SILVENTOINEN goes over environmental and genetic factors that had an impact on an adult's height. This paper helped me understand that each country may have a deeper reason for their average height. Which made me look for environmental or economical factors in countries that could have a correlation with height. One major con that they experienced is they had a major lack of data and their data collection methods were very weak, this reflected also onto my own research as finding data was challenging. Predicting women's height from their socioeconomic status: A machine learning approach, by Adel Daoud, Rockli Kim, and S.V. Subramanian. These researchers used ML so that it could increase relative to OLS performance. Which would be good in predicting height using datasets about low and middle income countries. But they faced the same problem about lack of data. But for that data and source that they did have, was used to help with my research.

Dataset

The dataset name is “COUNTRY-SPECIFIC DATA" by NCD Risk Factor Collaboration. The dataset that I was working with was a numerical dataset. “The computer code for the model is available online, as are our country and regional estimates both in numerical format”(NCD Risk Factor Collaboration) In the dataset there were 136,500 samples. For preprocessing the data I needed to drop repeated variables such as year, country named , and country code. Also there were other unneeded things in the dataset that also needed to be dropped such as mean height standard error, Mean height upper 95% uncertainty interval, and Mean height lower 95%

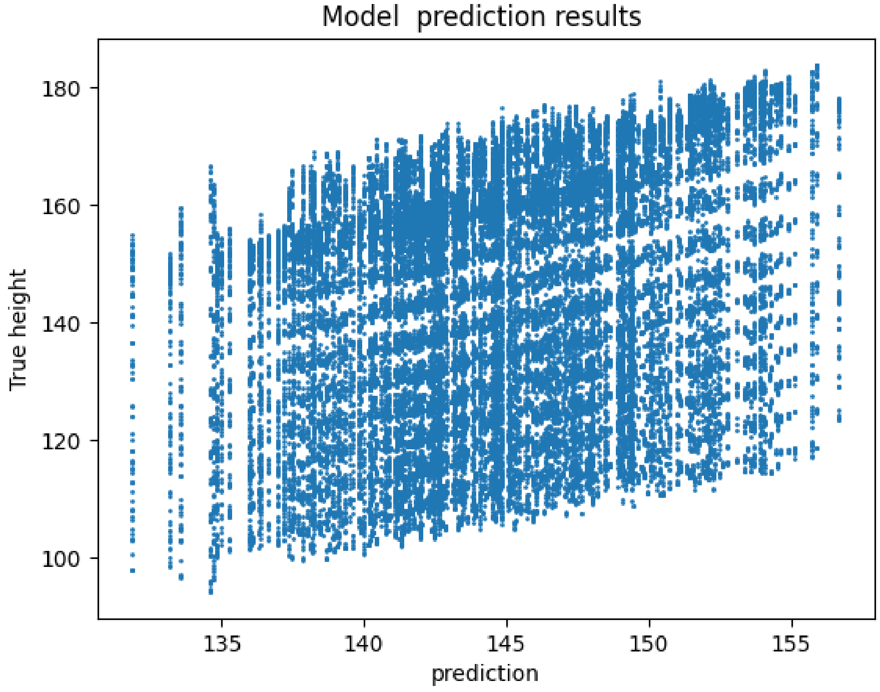
uncertainty interval. So the main things I was left with was mean height, age, and year. These needed to be dropped since it was a string being read as an integer. ([originally I had added the country gdp as one of my data sets which i combined this original dataset with. But then realized later that it was unneeded because there was a lack of other country similarity datasets which were primary and removed from my dataset.])



Methodology/Models

Decision tree regressor, linear regression, and random forest regressor were all used in order to predict height. The Decision tree regressor that I worked on works by breaking down my data set into smaller parts until it starts getting closer to the true value. Linear regression works by creating a line of fit and making educated guesses using the line. The random forest regressor works by using yes or no questions to narrow the result to a valid choice. use of these models caused a problem with my model, causing it to not work since some of my data included country name and gender. Since country name and gender were a string, and my model needed floats, and integers to operate. To combat this problem, I used “pd.get_dummies” to translate my data into a checkbox style dataset. Meaning whenever that country was used, or gender was used, it would input a number 1. If not used, it would put a 0 implying it should be skipped over. Dropped columns of the original dataset included year, sex, mean height lower 95% uncertainty interval, mean height upper 95% uncertainty interval, mean height standard error, and country. My Y value in my test model was mean height. While the X value was imputed as everything other than age group and mean height. Because of the creation of dummies, I could drop the age group column. I split the dataset into train and test using “train_test_split()”. And made the test size 0.2.

Results and Discussion



The model metrics used were Mean Absolute Error, Mean Squared absolute error, and R^2 score. Originally, using only Year, the Mean absolute error under performed with a score roughly around 16.93. Mean Squared absolute error resulted with a score of roughly 381.5958.

And r^2 performed with the resultant score of 0.0007534. After further understanding, columns, age group and country name were added. Resulting in a slight increase of results. Mean absolute error results performed slightly worse, with results of 17.09. Mean Squared absolute error improving to 363.36. R^2 performed 2 orders of magnitude better with the resultant of 0.0607772. The correlation between height, age group and country is there, but there should be more data that would correlate to height. (Louis Lello, Steven G Avery, Laurent Tellier, Ana I Vazquez, Gustavo de los Campos, Stephen D H Hsu) published a research paper that analyzed genomic predictors and how it could be used to predict human height. While (KARRI SILVENTOINEN) found “While variation in body height is mainly due to genetic factors, environmental factors also have a substantial effect”. But the lack of data that corresponds with the height prediction data did not line up, meaning we could not use this information to benefit the study fully. My model performed poorly because of the relationship between country, gender with mean height. Although data such as (Galtons Parent and Child Data) show correlation between parental height and children's height.

Conclusion

Linear regression, Decision Tree Regressor, and Random Forest regressor performed poorly. Most likely because of the lack of data variety, that may correlate to height. Each model was roughly around 17cm off the real height. The use of year, age, and mean height is not enough to accurately predict height so that it could be used in the medical world. Something I would have changed with this research work would be the fact that the height prediction was mainly for humans young of age. Something that would match well with height prediction, according to(Julia Schäppi, Silvia Stringhini,Idris Guessous, Kaspar Staub,corresponding author, and Katarina L Matthes) “Moreover, the effects of body height on health are still considered too little in the public health field. This is especially true for height loss in the second half of life, where better data are needed.” Meaning for further analysis on older people's height, the knowledge of height decay should be studied as well.

Acknowledgements

Would like to acknowledge Jonathan Lorenzo with helping guide my research. And Albert Stanley with advice about my research.

References:

Daoud A, Kim R, Subramanian SV. Predicting women's height from their socioeconomic status: A machine learning approach. Soc Sci Med. 2019 Oct;238:112486. doi: 10.1016/j.socscimed.2019.112486. Epub 2019 Aug 14. PMID: 31470245.

Louis Lello, Steven G Avery, Laurent Tellier, Ana I Vazquez, Gustavo de los Campos, Stephen D H Hsu, Accurate Genomic Prediction of Human Height, *Genetics*, Volume 210, Issue 2, 1 October 2018, Pages 477–497,

Silventoinen K. Determinants of variation in adult body height. J Biosoc Sci. 2003 Apr;35(2):263-85. doi: 10.1017/s0021932003002633. PMID: 12664962.

Akachi Y, Canning D. The height of women in Sub-Saharan Africa: the role of health, nutrition, and income in childhood. Ann Hum Biol. 2007 Jul-Aug;34(4):397-410. doi: 10.1080/03014460701452868. PMID: 17620149.

D. Rativa, B. J. T. Fernandes and A. Roque, "Height and Weight Estimation From Anthropometric Measurements Using Machine Learning Regressions," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 6, pp. 1-9, 2018, Art no. 4400209, doi: 10.1109/JTEHM.2018.2797983.

Height and Body-Mass Index Trajectories of School-Aged ... - the Lancet, www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31859-6/fulltext. Accessed 13 Nov. 2023.

“Hub: Determinants of Variation in Adult Body Height: 10.1017/S0021932003002633.” *Sci*, sci-hub.hkvisa.net/10.1017/s0021932003002633#google_vignette. Accessed 12 Nov. 2023.

Lello, Louis, et al. “Accurate Genomic Prediction of Human Height.” *OUP Academic*, Oxford University Press, 27 Aug. 2018, academic.oup.com/genetics/article/210/2/477/5931053.

Schäppi, Julia, et al. “Body Height in Adult Women and Men in a Cross-Sectional Population-Based Survey in Geneva: Temporal Trends, Association with General Health Status and Height Loss after Age 50.” *BMJ Open*, U.S. National Library of Medicine, 8 July 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC9272122/.

"Wolfram Data Repository." *Galton Parent and Child Height Data | Wolfram Data Repository*, 18 Nov. 2022, datarepository.wolframcloud.com/resources/Galton-Parent-and-Child-Height-Data/.