Data Competition Project
Advanced Machine Learning
Spring 2025
Due: Mar 2nd (leaderboard) and 3rd (report), 2025 at 11:59p

Project Description

You will work in randomly assigned groups of 3 to complete this project, consisting of two parts:

1. A Kaggle competition in which you build a model and predict on a test set of data, and
2. A written report which details your model building process and answers a few scientific questions relevant to the dataset, either with your Kaggle model or other statistical methods. ****Submit complete and documented code along with your report.****

Premise and Data Description

You are an alien scientist on a distant planet that is home to a species of lifeforms which are remarkably similar to humans in their physiology, reproduction, and development. Scientists on this planet have conducted a longitudinal study to research the growth process of aliens from the time of birth until they reach adulthood (around 18-20 years).

In the first generation of this study, the standing heights of subjects in Cohort 1 were recorded at several time points throughout their development, until around age 20.

In the second generation of this study, the progeny of the individuals in the Cohort 1 were followed from birth to around age 18. The standing heights and weights of the individuals in Cohort 2 were tracked.

Warning: This dataset has many complications, including but not limited to: *irregular longitudinal design, subject dropout, missing data, outliers, and potential data contamination.* Some Gen 1 subjects also may not have children who are present in Gen 2. Pay special attention to how you address these issues and document them in your written report.

Figure 1: A friendly alien and their offspring, participating in the longitudinal study.

There are two main modeling tasks that the scientists want to achieve:

1. *For the first task, you are asked to construct a machine learning model which aims to predict the growth curves of Cohort 2 aliens from ages 10 to 18, using their own recordings from age 0 to 9 and their parents complete growth curve, if available. You are given a training set which consists of complete data for a training set of Cohort 1 and Cohort 2 subjects, as well as predictor data for a test set of subjects, for whom you will need to predict the second half of growth curves from ages 10 to 18. Your predictions will be uploaded to Kaggle and evaluated on a leaderboard. (See the Kaggle page for more details on how to format your data for proper evaluation.)*

   You will also need to write up a description of your Kaggle model and document the model development process in the first part of your written report.

2. *In the second part of your written report, using either your Kaggle model or other statistical/machine learning methods, you must conduct an exploratory data analysis and quantitatively answer the following scientific questions about alien heredity in growth across generations:*
   a. What features of a parent's growth curve are most predictive (if at all) of a child's **final height at adulthood**?
   b. What features of a parent's growth curve are most predictive (if at all) of the **magnitude (amount grown)** of a child's pubertal growth spurt, typically occurring between the ages of 9 and 15)?
   c. What features of a parent's growth curve are most predictive (if at all) of the **timing (onset and duration)** of a child's pubertal growth spurt, typically occurring between the ages of 9 and 15)?
   d. For questions b. and c., does the strength of heredity change depending on the relationship between parent/child sexes assigned at birth? (E.g. parent/child combinations of matched sexes – M/M, F/F, or opposite sexes – M/F, F/M.) If yes, how do these associations change in each case?

   (Hint: for growth spurt analysis, you may want to consider the growth velocities instead of the raw height measurements.)

   Note: For the second task, you only need to use the training data, since the test data is censored for Gen 2 beyond age 9.

   Write up your analysis and responses to these questions in the second part of your report, bolstering your analysis with a combination of visualizations, statistical metrics, and hypothesis testing.

Evaluation

Teams will be evaluated according to the following rubric:

	Rubric Description	**Points**
Data Competition	*Based on the private leaderboard, scores will be assigned at each landmark as follows:*	Total: 30
Leaderboard standing at Checkpoint 1 (Feb 7 @ 5p)	Top quartile - 5 Mid 2 quartiles - 4 Bottom quartile - 2.5	/5
Leaderboard standing at Checkpoint 2 (Feb 28 @ 5p)	Top quartile - 5 Mid 2 quartiles - 4 Bottom quartile - 2.5	/5
Final leaderboard standing (Mar 2 @ 11:59p)	Top quartile - 20 Mid 2 quartiles - 15 Bottom quartile - 10	/20
Written Report (due Mar 3, 11:59p)	****Submit complete and documented code along with your report.****	Total: 60
Intro, Exploratory Data Analysis, & Data Cleaning	Accessible intro, helpful and insightful visualizations, and detailed description of how data problems (irregular longitudinal design, subject dropout, missing data, outliers, etc.) are addressed for model development and heredity analysis.	/8
Description and Development of Final Kaggle Model + Code Submission	Well-written and thorough description of your model, its architecture, and your model performance using evaluation metrics. Includes discussion of the model development process.	/20
Heredity Analysis:		
Question a. (Final height at adulthood)	Clear description of feature engineering and statistically sound analysis of hereditary association. Easy to read written conclusion which cogently answers the scientific question.	/8
Question b. (Growth spurt magnitude)	Clear description of feature engineering and statistically sound analysis of hereditary association. Easy to read written conclusion which cogently answers the scientific	/8

	question.	
Question c. (Growth spurt timing)	Clear description of feature engineering and statistically sound analysis of hereditary association. Easy to read written conclusion which cogently answers the scientific question.	/8
Question d. (Parent/child heredity across sex combinations)	Detailed and thorough comparison of heredity across parent/child sex combinations. Easy to read written conclusion which cogently answers the scientific question.	/8
Teamwork and Partner Evaluation		Total: 10
Collaboration Score (Peer Review)	Were you a good partner to work with?	/10

Bonus Opportunity

The top 5 teams on the private leaderboard will have the option (*but are not required*) to present their final models in a slide deck (approx. ~5-10 min per group) on the last day of class for up to 5 bonus points on their project grade.