Logbook Details

Science Fair Logbook 2026

Wednesday, January 7, 2026
Time Spent: 2 hours
Location: Home

Goals for Today:
- Set up technical environment
- Find a list of potential datasets
- Identify knowledge gaps through reading article

What I did:
- Set up Colab and performed a test run to ensure that it works
- Found 2 datasets and listed info, pros, and cons of each (See Data Collection for more)
- SUMMARY OF KNOWLEDGE GAPS

Challenges:
- Finding a data set that was suitable in all ways for my project

Next Steps:
- Tomorrow: finalize the dataset and start learning Python basics

Questions for teacher/mentor?

Resources Used:
- kaggle.com
- Google colab
- Google Dataset Search
- My Medium article

Wednesday, Jan 28th
Time Spent: 2 hours
Location: home

Goals for Today:
- ☑ ~~Learn and practice Python Basics~~
- ☑ ~~Learn pandas basics~~

- Watched youtube videos to understand fundamental concepts
- Replicated the code on youtube to practice fundamentals

Challenge:

- Many roadblocks when learning. Finding and loading dataset onto google colab and imprint pandas (virtual environment setup)

Resources:
- Programming with mosh “Learn Python in 1 Hour”
- 30 Min Pandas Tutorial: Learn Pandas in 30 Minutes - Python Pandas Tutorial

Thursday, Jan 29th
Time Spent: 1 hour
Location: Home

Goals for today:
- ☑ ~~Learn matplotlib~~
- ☐ find ML resources

Challenge:
- Hard to know which concepts are important to grasp for the project

Resources:
- Gemini to help me find videos that teach me the ML basics that I need to know

Sunday, Feb 1st
Time Spent: 2.5 hours
Location: home

Goals for Today:
- ☑ ~~Finish watching ALL ML videos~~
- ☑ ~~Gain understanding of ML and replicate some basics~~

Challenge:
- Too much info, hard to know which info is important
- Straightforward day, little to no challenges

Resources:
- Start using Matplotlib in 7 minutes! 📊
- Statquest ML basics playlist on Youtube (SUPER HELPFUL)

Monday, Feb 2nd
Time spent: 3 hours
Location: home

Goals for Today:
- ☑ ~~Set up digital environment~~

- [x] ~~Find dataset~~
- [ ] Clean and process data set

- Found this dataset: https://archive.ics.uci.edu/dataset/565/bone+marrow+transplant+children
- Other datasets that I am not choosing:
- https://data.mendeley.com/datasets/myd48fczg4/1
- https://cibmtr.org/CIBMTR/Resources/Publicly-Available-Datasets#
- Chose the first data set because: size, cleaness, ability, relevance to my project

Set up google colab
- Import the dataset (HARD and CONFUSING)
- Load in pandas

- Did not get to dataset cleaning

Challenges:
- Google workspace kept giving me errors and the code wouldn't run
- The dataset I tried to import needed unzipping and was an arff file UNTIL I uploaded it to drive
- I needed to mount drive which was hard
- Less motivation as time was crunched, one setback was mentally tolling

Resources:
- Google colab
- Gemini to help me debug
- Dataset: https://archive.ics.uci.edu/dataset/565/bone+marrow+transplant+children

Tuesday, Feb 3rd
Goals:
- [x] ~~Clean and filter data~~
- [x] ~~Follow ML regression tutorial to build model~~
- [x] ~~Relate project to Biology – get Biology data and information and research~~

What I did:
- Scaled and imported NaN on the data
- Did a LOT of debugging – trying to figure out why the code wouldn't run, issues with google dring mounting
- Followed full ML tutorial to build my model – Build your first machine learning model in Python
- Analysed the data shown after building the model

Challenge:

- Random forest model has a NEGATIVVE R2 because I used the scaled data. I switched scaled data to imputed data and my random forest model improved drastically

- Tailored the video to my needs
- Lots of DEBUGGING in many parts of my code

Resources:
- Gemini for debugging
- https://www.youtube.com/watch?v=29ZQ3TDGgRQ - ML tutorial

Wednesday, Feb 4th 2026
Time spent: 4 hours
Location: Home
Goals for Today:
- ☑ ~~Finish writing project info on CYSF~~
- ☑ ~~Finalize and download Logbook~~
- ☑ ~~Final Code Debug~~
- ☑ ~~Logbook entry~~

Challenges:
- Volume of work and time management

Next steps:
- Finalize pitch and presentation
- Set up trifold

Resources:
- Google colab
- Google docs

Thursday, Feb 5th 2026
Goals for Today:
- ☑ ~~Finalize pitch and presentation~~
- ☑ ~~Set up the trifold~~
- ☐ Make a list of resources (youtube, colab, gemini, etc.)

Challenges
- Printer only available at school

Next Steps:
- School Science Fair

Resources:

- Trifold, art supplies, printer
- Google docs
- Google colab

Things I want to talk about during pitch

- Regression vs classification choice
- Threshold of when to alert the doctors
- How I improved the random forest by changing scaled* to imputed* — went from negative r2 to very good r2
- The table of values of how well each one did
- The one significant outlier
- Medical impicaltions and how it applies (ask chat)

Background Research

Feature Definitions

Feature Name	Definition
aGvHDIIIV	The Target. Whether the patient got severe (Grade II-IV) Acute Graft-versus-Host Disease.
CD34kgx10d6	The number of Stem Cells given to the patient per kg of their weight. This is a crucial dosage metric.
CD3dCD34	The ratio of T-cells to Stem cells in the transplant.
Recipientage	The age of the child receiving the transplant.
Donorage	The age of the person who donated the marrow/cells.
Rbodymass	The Body Mass Index (BMI) or weight of the recipient.
PLTrecovery	Platelet Recovery. How long it took for the patient's blood to start clotting on its own again.
ANCrecovery	Neutrophil Recovery. How long it took for the patient's immune system (white blood cells) to start working.
survival_status	Whether the patient survived the follow-up period (0 = Alive, 1 = Deceased).

TITLE HERE

- Four stages: mild, moderate, severe, life-threatening
- aGvHD occurs when the stem cells in the bone marrow recognize your child as different from what they expect
- Bone marrow synthesizes red blood cells
- Acute GvHD develops early after the transplant (less than 100 days)
- Graft (donor) contains white blood cells called lymphocytes that can tell whether or not a cell belongs in your body.
- GvHD – graft sends out lymphocytes to attack the child's organs and tissues
- Different treatments have been made to treat HCT (bone marrow transplant)
- Autologous HCT or Allogeneic HCT
- Allogeneic HCT takes healthy donor stem cells that are matched
- Autologous HCT uses patients own stem cells, often to recover from high dose of chemotherapy
- HCT restores normal blood cell production (hematopoiesis) in the bone marrow

Write Up

Abstract
Acute graft-versus-host disease occurs when the immune cells from the donor attack the patient’s tissues. This happens within 100 days of the allogeneic transplant. Using an ML regression model, we can predict a patient's severity score given sufficient data. In the dataset used, children who have undergone a Hematopoietic Cell Transplant (HCT or Bone Marrow Transplant) have been studied and monitored for GvHD. This project demonstrates the use of a Regressive Machine Learning Model to score a patient on GvHD risk.

Problem
In order to enhance and improve our current treatments for HCT, we need to build something fast, but also safe. In the past, checking patients for GvHD risk required many pre-clinical safety screenings. These screenings often took weeks of testing. To make the treatment even faster than it was before, the next steps are to integrate Applied Machine Learning to predict whether or not a patient has a high or low risk of developing GvHD. By predicting risks ahead of time, many lives can be saved. ML models can rapidly filter out patients who are unsuitable for the treatment, speeding up the process significantly. In a clinical setting, doctors may only be able to test a few factors to determine whether or not a patient is considered safe for HCT treatment. Having an ML model can analyse many factors at once and come up with a decision for doctors to then double-check.

Method
I started by finding suitable datasets for my project by looking at public and clean datasets. I used a GvHD prediction dataset named “Bone Marrow Transplant: Children”. Using Google Colab, I loaded the data into Google Colab. I then set up my digital workspace, which included Pandas, Matplotlib, Scipy, and more. The dataset had missing values and needed cleaning. I started off by replacing all values that were missing with NaN (Not a Number). Then, I replaced any typos with NaN. For example, if someone had put 25yo as patient age instead of 25, the model would replace it with NaN. This ensures that the model faces little to no errors. Next, I got to know the data better by using df.head() and analysing the data. I plotted a feature correlation map with the target row being aGvHDIIIIV. By looking at the target row, I could tell which features were more related to each other and which ones had little to no relation. To read this chart, you must check to see which boxes are brightly colored along your target row. Then I removed my “cheater rows,” which are rows that happen after the doctor checks to see if GvHD has occurred. By removing these rows, we are lowering the accuracy of the model, but we are making the model more true to clinical settings. It also significantly prevents overfitting, which is something that I struggled with when building. Then, I split the data into 80% training data and 20% testing data. The split is crucial in ML as the model can not be trained on data it has already seen. I implemented the median strategy on the dataset, which uses the median of some values to avoid having many empty parts of the data. I also scaled the data for only the linear regression model, as scaling does not work well for random forests. I trained the model to make a prediction using y_lr_train_pred and evaluated the performance of the linear regression model and the random forest model using R2 and MSE. Finally, I compared the values of both models and found out that the random forest works better in this case. I created a plot with a trend line and jitter to present the accuracy of my model.

Analysis

Initially, my results were overfit, which led to the training data being more fit than the testing data. This essentially means that the Machine Learning Model memorized the data and just recited it for the test. When predicting, the model relied solely on the time_to_aGvHD_III_IV feature, meaning that it was taking shortcuts. To fix the overfit, I cut three features that were allowing the model to cheat. I believe that the Machine Learning Model that I built is a good baseline product that can be expanded and made more complex in order to achieve higher accuracy. The linear regression model failed with a negative R2. This means that biology is complex and not built linearly. In this case, the random forest worked much better. While 15.5% R2 may be low in a laboratory setting, it is a considerable amount considering the real-world aspects and fluctuations of this dataset. There are also outliers in my dataset and graph. It is a false negative, which I believe is due to unnatural biological circumstances.

Conclusion

ML models can assist doctors and healthcare professionals in a clinical setting in decision-making. They can significantly decrease the time needed to assess whether or not a patient is suitable for a treatment and help limit human mistakes. In the future, I want to apply a threshold for when to tell doctors a positive or negative result. For example, I can set a value of 0.5, and anything over 0..5 would be considered severe. This way, patients can get a severity score for GvHD and also a classification output. Other ways to improve the model are to filter out features with low impact or with many missing values. Or use K-Fold Cross-Validation to improve accuracy and reduce overfitting.

Citations

AboutKidsHealth. (2010, March 6). *Maladie du greffon contre l'hôte après une greffe de sang et de moelle osseuse*. https://www.aboutkidshealth.ca/fr/santeaz/haematology/maladie-du-greffon-contre-lhote-apres-une-greffe-de-sang-et-de-moelle-osseuse/?language=en#:~:text=Key%20points,it%20is%20called%20acute%20GVHD.

Bone Marrow Transplant Acute Graft vs. Host Disease. (n.d.). https://www.nationwidechildrens.org/family-resources-education/health-wellness-and-safety-resources/helping-hands/bone-marrow-transplant-acute-graft-vs-host-disease#:~:text=After%20someone%20has%20a%20bone,someone%20other%20than%20the%20patient

Data Professor. (2022, May 27). *Build your first machine learning model in Python* [Video]. YouTube. https://www.youtube.com/watch?v=29ZQ3TDGgRQ

Goldsmith, S. R., Ghobadi, A., Dipersio, J. F., Hill, B., Shadman, M., & Jain, T. (2022). Chimeric Antigen Receptor T Cell Therapy versus Hematopoietic Stem Cell Transplantation: An Evolving Perspective. *Transplantation and Cellular Therapy*, *28*(11), 727–736. https://doi.org/10.1016/j.jtct.2022.07.015

Tech With Tim. (2025, August 6). *Learn Pandas in 30 minutes - Python Pandas tutorial* [Video]. YouTube. https://www.youtube.com/watch?v=EXIgjIBu4EU

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/565/bone+marrow+transplant+children

Acknowledgement

BLANK

Presentation

Attachments

- Logbook
- Google Colab

Declarations

Everything is my work

Final Findings

Outcome Graphs and Tables are Embedded in the Code

R2 and MSE Values (Linear Regression and Random Forest)

```
[31]
✓ 0s
```

```
df_models.reset_index(drop=True)
```

	Method	Training MSE	Training R2	Test MSE	Test R2
0	Linear regression	0.00485	0.971243	0.03458	0.791946
1	Random forest	0.000435	0.997421	0.017695	0.893537

- The above is my overfit data, which I realised **because the training MSE is much better than the testing MSE**
- Random forest has a lower MSE and a higher R2
- I realized that random forest does not work well with scaled data. **I changed to imputed data, and R2 and MSE improved significantly**
- R2 was negative, which means that the model is doing worse than if you randomly selected values
- R-squared: the proportion of variance in the dependent variable by the independent variable (squared so that all is positive)
- Mean Squared Error: average of the squared differences between actual and predicted values

Important Features

- Heatmap shows that PLTrecovery (platelet recovery: 0.35), survival status (whether or not the patient survived: 0.35), and Rbodymass (BMI or weight of participant, 0.24)
- Jitter plots show severity and allow the graph to be more visually appealing

Medical Application

- False negatives (when the model says the patient is safe, but they actually have GvHD) are the most dangerous
- The Future: **threshold applied to determine when to tell doctors** a positive GvHD (ex. >0.5 severity)

Overfitting

- Why it may be overfit: small sample size, max depth not shallow enough, too many missing values
- In the future, I would **use Feature Selection to remove features** with little to no impact on the final result. I would also **implement K-Fold Cross-Validation**
- Some features, like **extcGvHD**, have **31 missing values** and can be removed

- K-Fold Cross-Validation: The data is split into 5 or 10 folds. The training data is 4 folds, and one fold is used for testing. The cycle is repeated until every patient has been a part of the testing set once
- To fix my problem, I **dropped the cheating variables that were recorded after GvHD was found**. time_to_aGvHD_III_IV, survival_time, and survival_status

Pitch

The Hook: A High-Stakes Medical Challenge

- **The Problem:** Acute Graft-versus-Host Disease (aGvHD) is a life-threatening complication in pediatric bone marrow transplants where the new immune system attacks the child's body.
- **The Goal:** I developed a machine learning pipeline to predict the risk of severe GvHD (Grades II-IV) using only data available on the day of the transplant.

The Methodology: Beyond Simple Math

- **Data Foundation:** I analyzed a dataset of 187 pediatric patients, focusing on pre-transplant "Day Zero" metrics like **CD34+ stem cell dosage**, **recipient body mass**, and **donor/recipient age**.
- **Model Comparison:** I compared two distinct algorithmic approaches: **Linear Regression**, which looks for straight-line relationships, and **Random Forest**, which uses an ensemble of decision trees to capture complex, non-linear biological patterns.
- **Preprocessing:** To handle real-world clinical gaps, I used a **SimpleImputer** to fill missing values and **StandardScaler** to ensure different units (like age vs. cell count) were weighted equally by the models.

The Pivot: Identifying "Data Leakage"

- **The Discovery:** My initial model showed a near-perfect **0.89 Test R^{2}**, but Feature Importance revealed it was "cheating" by using post-transplant data like time_to_aGvHD.
- **The Correction:** I performed rigorous **Feature Selection**, removing all "leaky" variables to ensure the model was performing a true prediction of the future, not just summarizing the past.

The Results: Honest Clinical Insights

- **The "Winner":** The **Random Forest** outperformed Linear Regression, achieving a **Test MSE of 0.140** and a **Test R^{2} of 0.155**.
- **The Reality Check:** While 15.5% may seem modest, it represents a genuine mathematical "signal" in a highly complex biological system where Linear Regression completely failed (Test R^{2} of -0.548).
- **Key Predictors:** The model identified that **platelet recovery time** and **stem cell dosage** are among the most significant early indicators of GvHD risk.

The Impact: AI as a Clinical Tool

- **Closing Argument:** This project proves that while AI cannot replace doctors, it can act as a "smoke detector." By identifying patients with a **high severity score** early, doctors can implement more aggressive monitoring or preventative treatments to save lives.