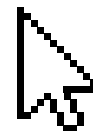




Machine Learning Applied in Crop Yield Prediction

2025/02/11 Tu B'Shvat

By Eric Gao & Mr. Pham





Scenario

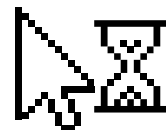


Bob is a farmer from Australia... The year is 2013

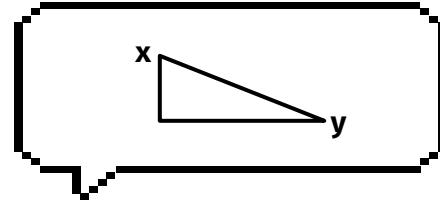
He wants to predict how much wheat his farm will produce this year in order to plan his financials and adjust his storage capacity

This is Bob's first year of managing his farm, so he found other farm's data across the globe in order to make a better decision

Help Bob predict how much wheat his farm will produce




<div><div>X</div></div>							<div>?</div>	
	#	Country	Item	Year	Average rainfall [mm]	pesticides used per tonne	Avg. Temp [C]	Area Yield [hg/ha]
	1	Australia	Wheat	2008	474.0	42935.38	16.7	10788
	2	Australia	Soybean	2012	476.0	48687.88	16.84	22598
	3	Austria	Wheat	2009	1110.0	3531.8	9.48	49295
	4	Armenia	Wheat	2007	562.0	241.71	9.44	25837
	5	Azerbaijan	Wheat	1999	544.0	148.68	12.74	19992
	6	Argentina	Wheat	1992	624.0	26156.0	17.51	21809
	7	Argentina	Soybean	2011	675.0	100424.62	17.58	26053
	8	New Zealand	Wheat	2010	1732.0	5086.0	13.54	81241
	9	Canada	Wheat	2007	537.0	45140.03	9.14	23317
	10	India	Wheat	2013	1083.0	45620.0	26.71	31538
	X	Australia	Wheat	2013	534.0	45177.18	17.4	?????



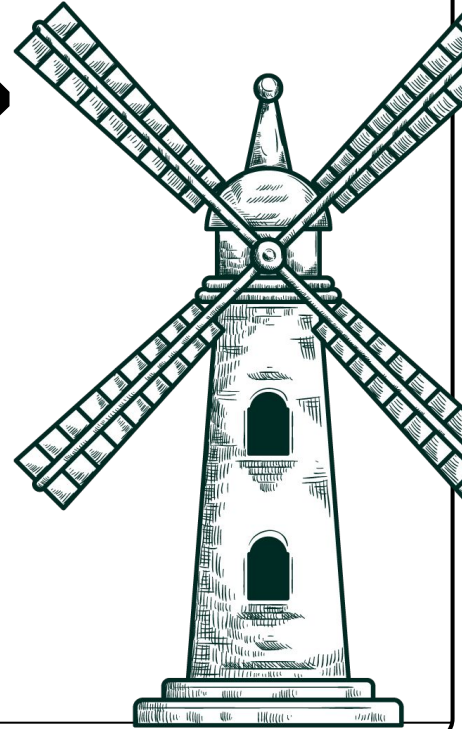
What are some **problems**
you encountered?

What are the problems?

- Multiple Influences
 - Interaction between the different effects
 - Regional Variability
 - Lack of understanding of ideal growing conditions for Wheat
 - Climate Change trends
 - Year to Year variability
 - Inconsistent records/data
 - Lack of concrete models
 - Data measurement standard difference
 - etc.
- 
- A solid green, wavy-edged shape located in the bottom right corner of the slide, resembling a stylized hill or a decorative graphic element.

What's the solution?

—Machine learning





01

The Problem

A deeper dive into the crop yield
prediction problem



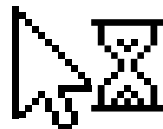
Facts



Overarching facts:

WHO estimate: 820 million people worldwide still have insufficient access to food

Food and Agriculture Organization (FAO) has projected a 60 percent increase in the demand for food to meet the needs of the projected global population of 9.3 billion by 2050





Why Crop Yield Prediction

Population

Optimize the allocation of resources such as water, fertilizer, and pesticides

Maximizing yields and minimizing input costs



Farmers

Implement risk management strategies to mitigate the negative impact

Stakeholders

Make informed decisions regarding market planning, trade policies, and commodity pricing

Scientists

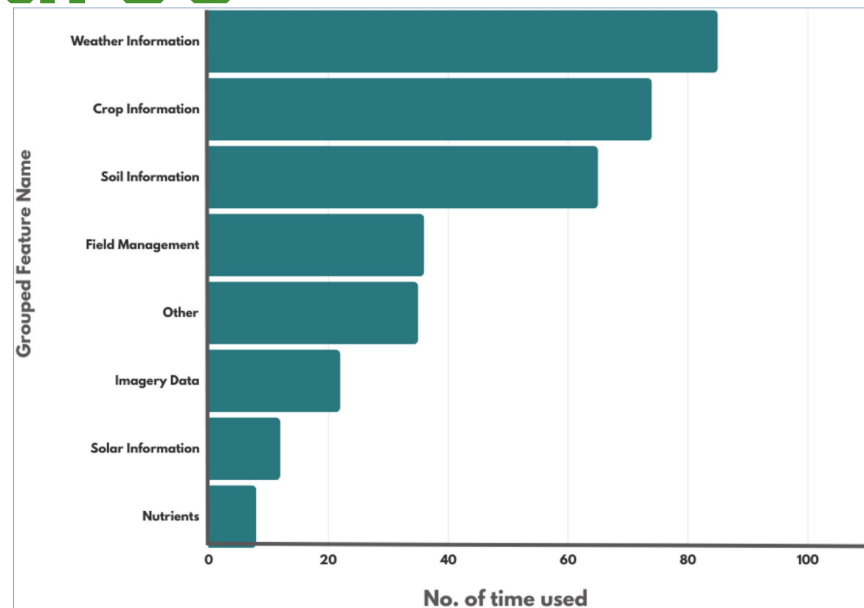
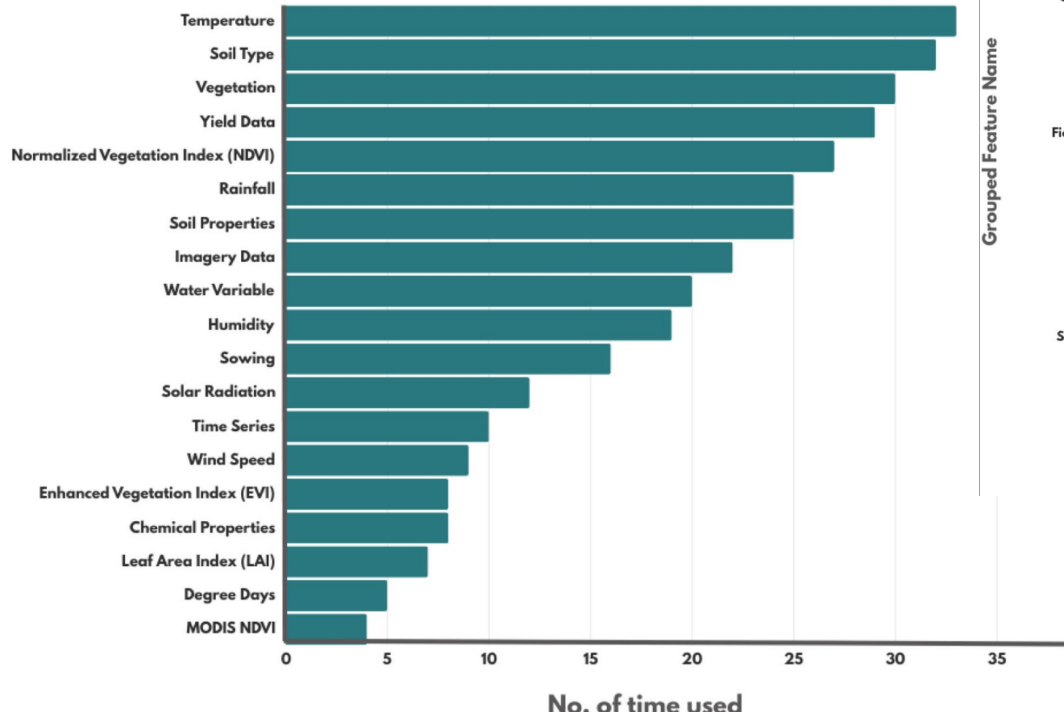
fundamental research question in plant biology, which is to understand how plant phenotype

Law makers

Develop evidence-based policies and programs aimed at promoting food security, sustainable agriculture



Most used Features



Crop yield prediction using machine learning: An extensive and systematic literature review

Sarowar Marshad Shawon ^o, Falguny Barua Ema ^o, Asura Khanom Mahi ^o, Fahima Lakman Niha ^o, H.T. Zubair ^o



02

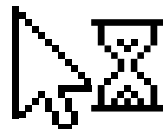
Machine Learning

Why Machine Learning?



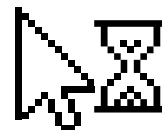
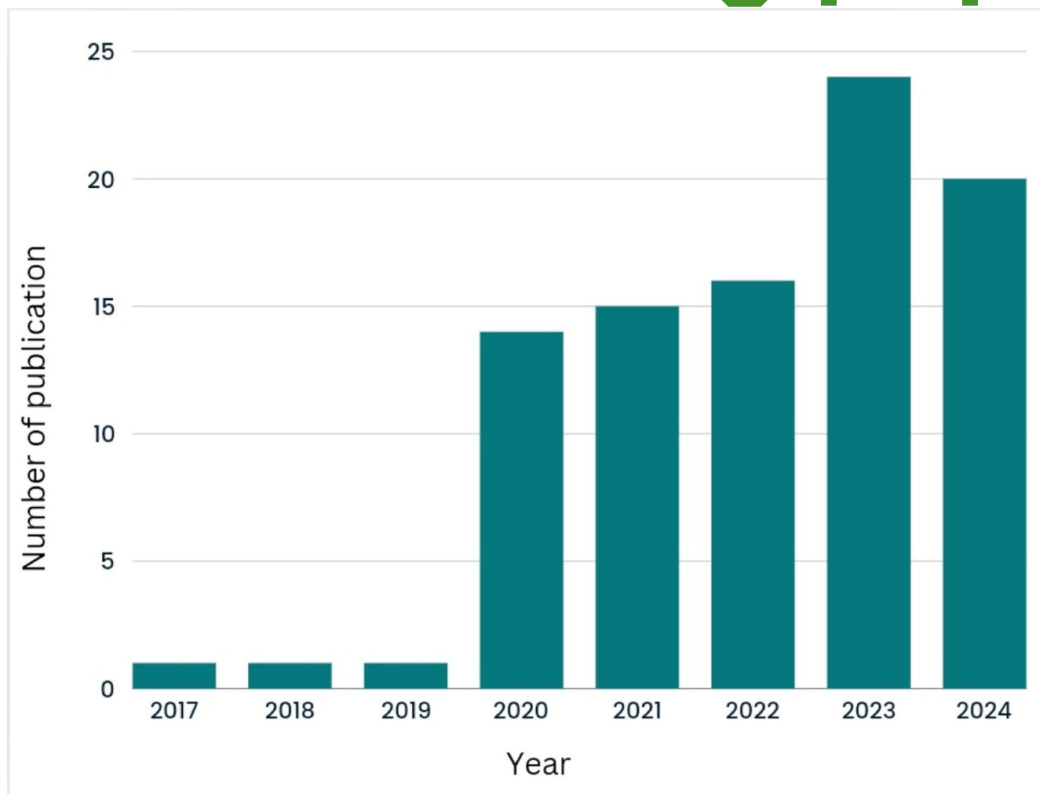
Machine Learning

- **Captures Complex, Non-Linear Relationships:**
 - Detect and model non-linear interactions
 - between rainfall and pesticide use
- **Handles High Dimensionality and Multicollinearity**
 - high intercorrelations among two or more independent variables
- **Robustness to Noise and Missing Data**
 - Built-in data imputation and are less sensitive to noisy or incomplete historical records





Machine Learning papers





Methodologies

Most used machine learning algorithms

No. of Time Used

Linear Regression	37
Random Forest (RF)	34
Gradient Boosting Trees	25
SVM (Support Vector Machine)	24
Decision Tree (DT)	15
k-Nearest Neighbor (k-NN)	12
Neural Network	9
Ensemble Learning	8
Logistic Regression	3

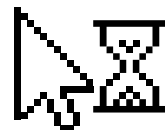


Metrics

Evaluation Metric	Used Number of Time
Root Mean Square Error (RMSE)	48
R-squared (R^2)	35
Mean Absolute Error (MAE)	31
Mean Squared Error (MSE)	12
Relative Root Mean Square Error (rRMSE)	6
Mean Absolute Percentage Error (MAPE)	6
Normalized Root Mean Square Error (NRMSE)	5
Median Absolute Error (MedAE)	2
Mean Squared Logarithmic Error (MSLE)	2

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - x_0)^2}$$

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$





03

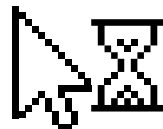
Dataset

A deeper dive



Dataset

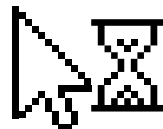
- From **FAO** (Food and Agriculture Organization) and **World Data Bank**
- **average_rain_fall_mm_per_year**: The average rainfall per year is approximately **1149**, with the least rainfall being 51 and the most 3240.
- **pesticides_tonnes**: The mean pesticides used in tonnes is **37077** tonnes, with minimum as little as 0.04 and maximum as huge as 367778 tonnes.
- **hg/ha_yield**: The average crop production yield is **77053.3**. Ranging from 50 all the way to 501412 hectograms per hectare.
- **101** unique countries/areas
- **10** unique types of crops are present





Dataset Train Test Split

- Partitioning dataset into two main subsets:
one for training the model and one for testing its performance
 - `train_test_split()`
- 80% for Training, 20% for Testing
 - Training Set: Used to “teach” the model to learn patterns.
 - Test Set: Held out and used only to evaluate how well the model generalizes to unseen data.
- Benefits:
 - Avoid a model that merely memorizes the training examples
 - Estimate Generalization Performance: Provides an unbiased estimate of how the model will perform in real-world scenarios
 - Allows for fair comparisons between different models





04

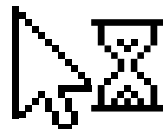
Level 1: Linear Regression

The sad algorithm that keeps being
compared to



Linear Regression

- Used **37** times total in peer-reviewed papers
 - Starting point of more advanced machine learning methods
 - Used as the “**control**” to compare other algorithms accuracy with
- Goal: Find a **straight line** that best fits a set of data points, so we can predict a continuous outcome
 - $Y = mx + b$





0.7486

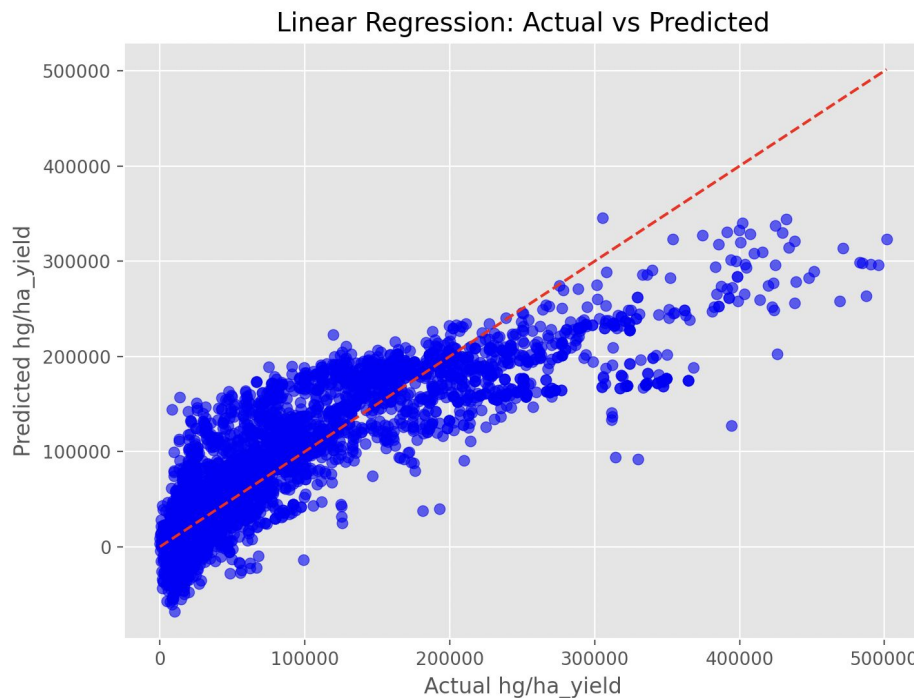
R-Squared Score

42681.52

Root Mean Squared Error

81051

Crop Yield Prediction for Bob
(Hectogram per Hectare)





05

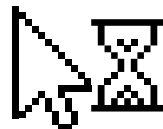
Level 2: KNN

Simple model that is surprisingly
accurate



K nearest neighbors (KNN)

- Used **12** times total in peer-reviewed papers
- Predicts the outcome for a new data point (e.g., the crop yield of a field) by looking at the 'k' most similar historical data points (neighbors) based on features such as, rainfall, temperature, etc.
- Vulnerability to **Noisy Data and Outliers**
- Choosing the right k (**k=2**)
 - Too few neighbors can lead to predictions that are overly sensitive to noise
 - Too many neighbors may smooth out important local variations





0.9859

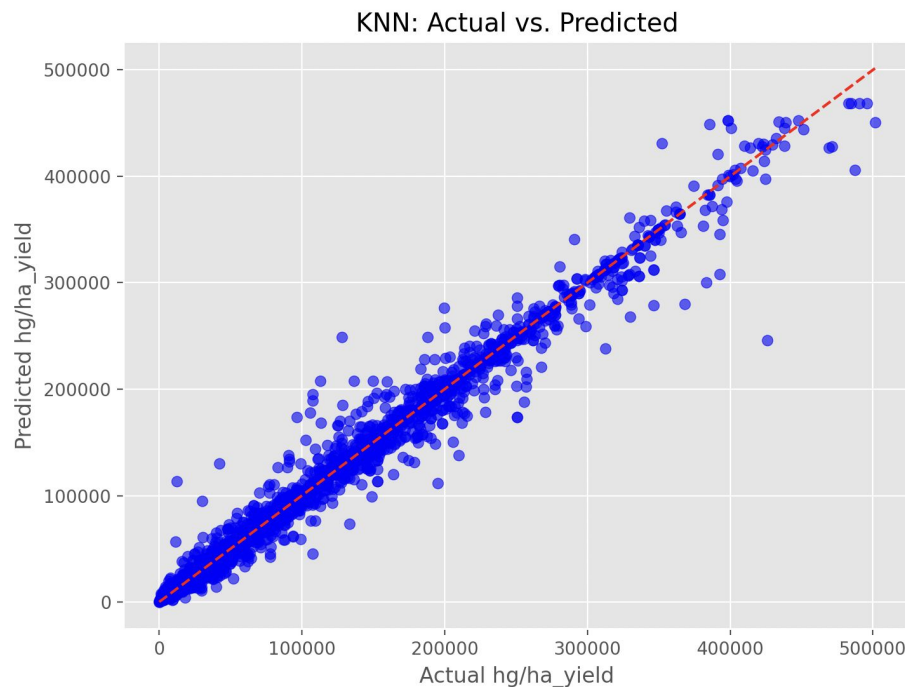
R-Squared Score

10090.16

Root Mean Squared Error

19560

Crop Yield Prediction for Bob
(Hectogram per Hectare)





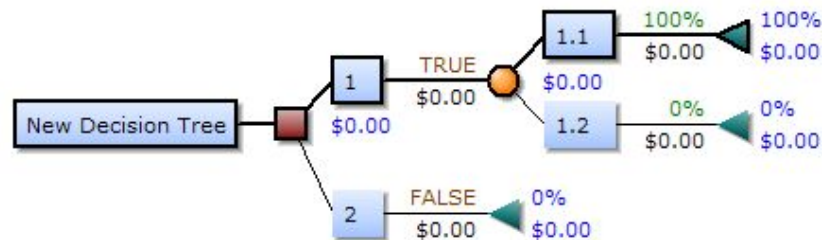
06

Level 3: Decision Tree

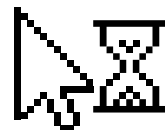
Tree...



Decision Tree



- Used **15** times total in peer-reviewed papers
- Splits historical agricultural data into branches based on **decision rules**
- Each branch represents a decision based on a specific feature, and the leaves (end nodes) represent the final crop yield predictions.
 - For example, a branch might split data based on whether rainfall is above or below a certain threshold, leading to different yield estimates
- Biased Towards dominant features
- Instability - Small change could lead to variation
- Overfitting Risk - Overly complex





0.9772

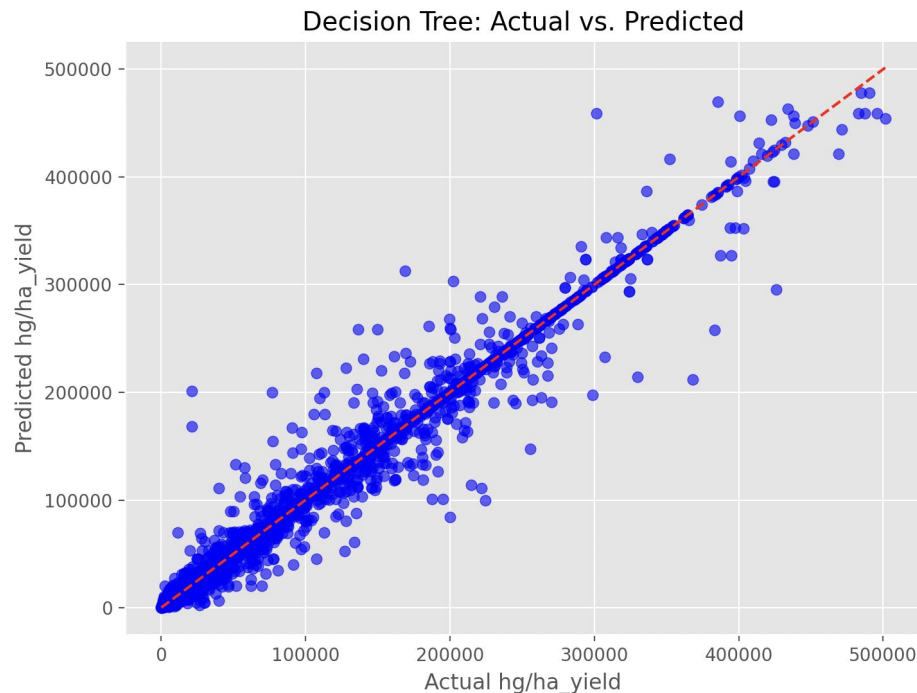
R-Squared Score

12855.79

Root Mean Squared Error

20301

Crop Yield Prediction for Bob
(Hectogram per Hectare)





07

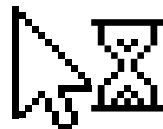
Level 4: Random Forest

Many Trees...



Random Forest

- Used **34** times total in peer-reviewed papers
- Uses historical agricultural data to grow **several decision trees**, each built on a random subset of data and features
- Each tree makes a prediction about crop yield, and the Random Forest aggregates these predictions (typically by averaging in regression tasks) to provide a final yield estimate.
- This ensemble approach helps capture various patterns in the data, improving overall prediction accuracy.
- Resource Intensive (My Laptop isn't good enough)
- Hyperparameter Tuning (I don't have enough time)





0.9795

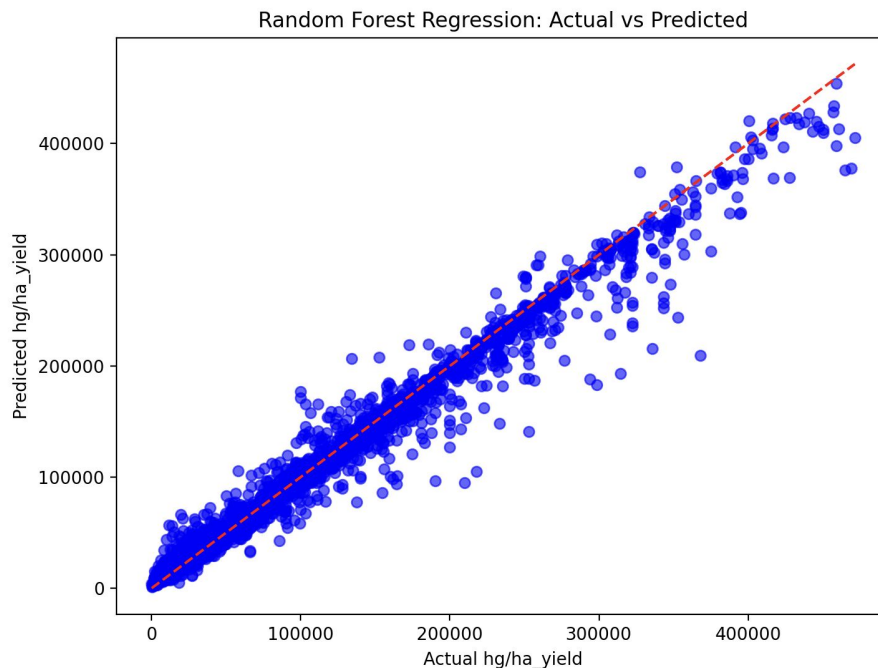
R-Squared Score

12050.96

Root Mean Squared Error

21697

Crop Yield Prediction for Bob
(Hectogram per Hectare)

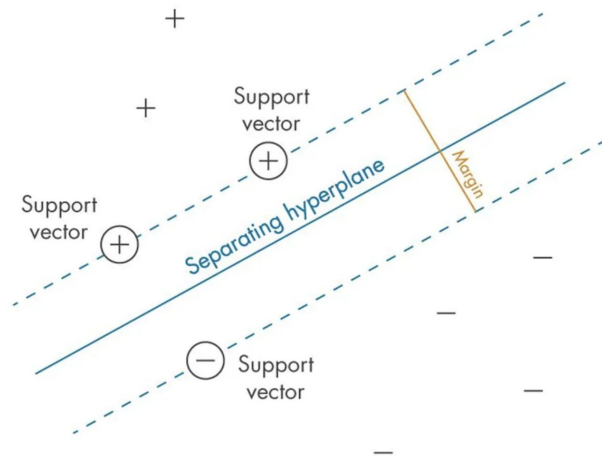




08

Level 5: SVM

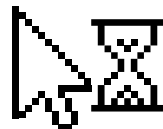
It took me 20 minutes to run this

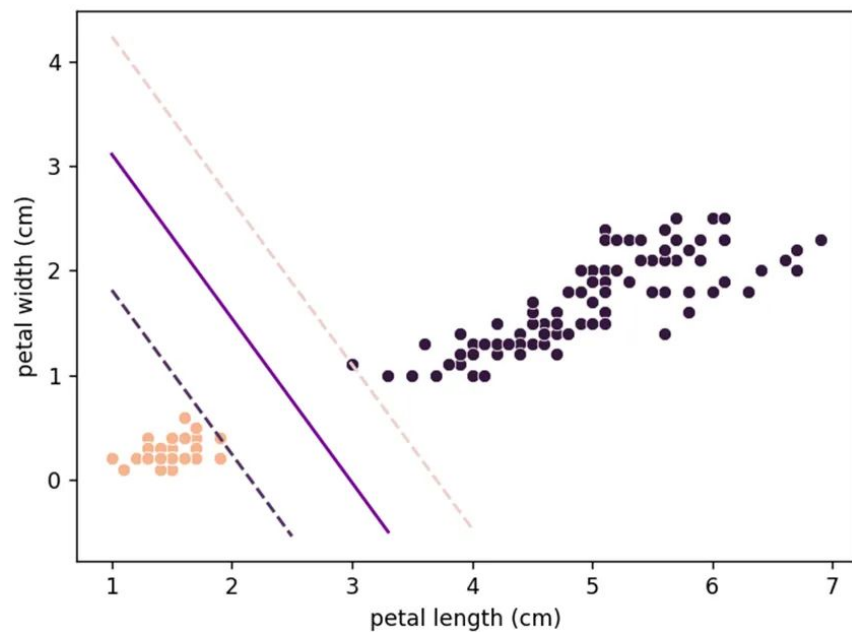
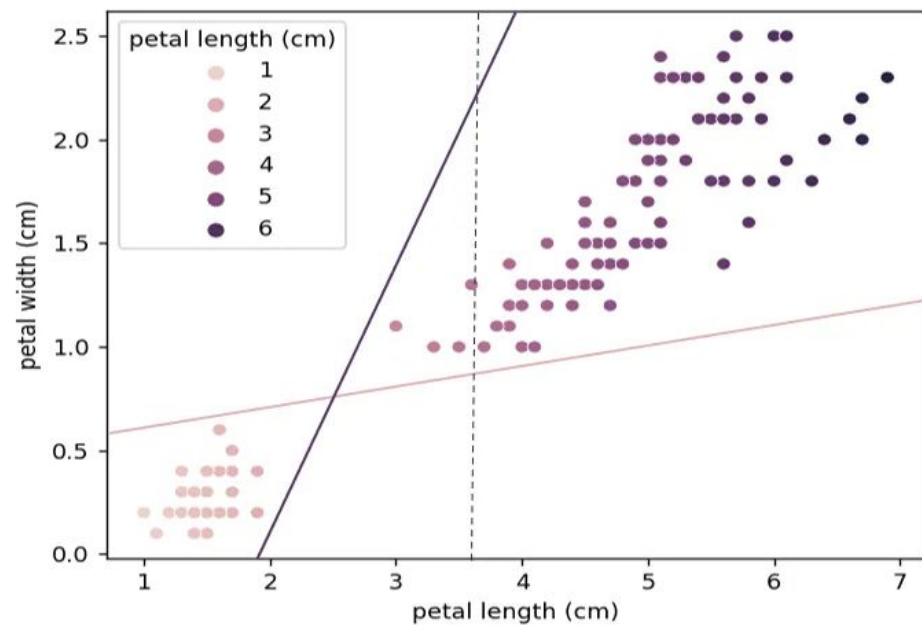




Single Vector Machine

- Used **24** times total in peer-reviewed papers
- It works by finding the optimal **hyperplane** (or boundary) that separates data points into different classes
 - Find position and orientation of the hyperplane to maximize the **margin** (greater degree of confidence)
- SVM seeks a function that maps input features (e.g., rainfall, temperature) to yield values
- Use the "kernel trick" to transform input data into a higher-dimensional space, enabling it to model non-linear relationships between agricultural factors and crop yield
- Computationally Intensive
- Sensitive to Parameter Tuning
- Limited Interpretability - Black Box







0.5565

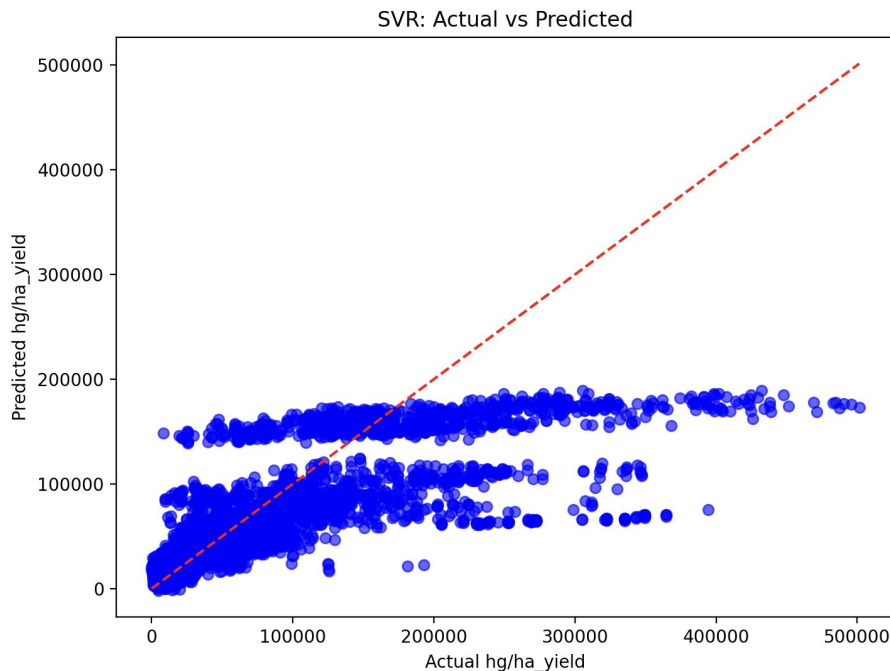
R-Squared Score

56691.26

Root Mean Squared Error

81360.98

Crop Yield Prediction for Bob
(Hectogram per Hectare)





09

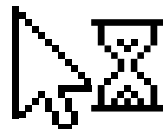
Level 6: Gradient Boosting Tree

It took me 20 minutes to run this



Gradient Boosting Tree

- **Sequential Learning:**
 - The process starts with a decision tree that makes initial crop yield predictions
 - Subsequent models focus on predicting the residuals (errors) of the previous model, gradually refining the overall prediction.
- **Aggregated Predictions:**
 - The final crop yield prediction is obtained by combining the outputs of all the models, resulting in a more accurate and robust forecast that can capture complex interactions in agricultural data.
- Computational Intensity
- Sensitivity to Parameter Tuning
- Risk of Overfitting (Noisy Data)





0.9694

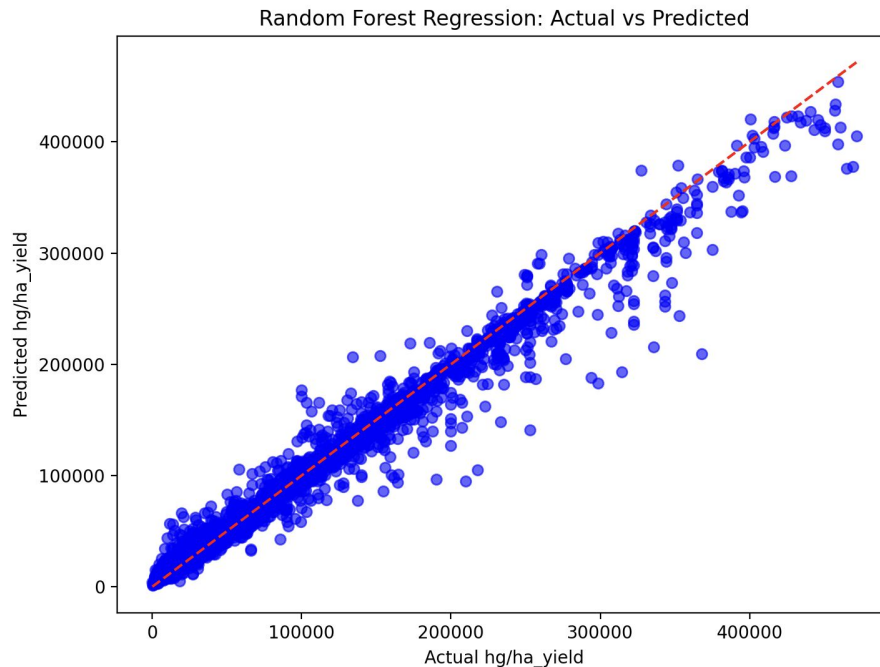
R-Squared Score

16471.29

Root Mean Squared Error

17599

Crop Yield Prediction for Bob
(Hectogram per Hectare)





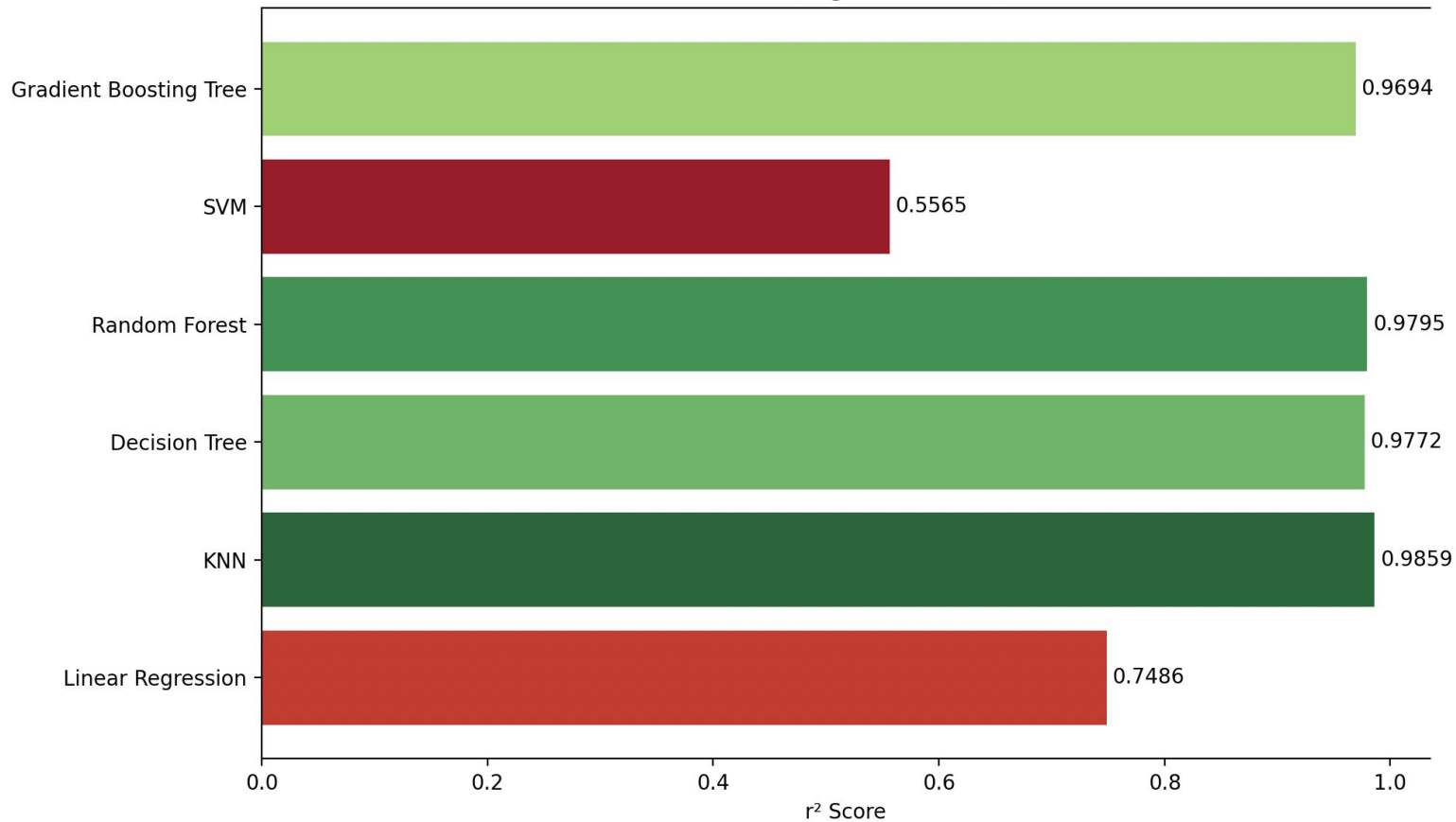
10

Results

Here we go



Model r^2 (Higher is Better)



Theoretical Ranking

Gradient Boosting Tree



SVM



Random Forest



Decision Tree



KNN



Linear Regression



Theoretical Rank (Lower is Better)

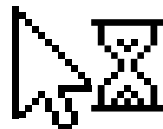


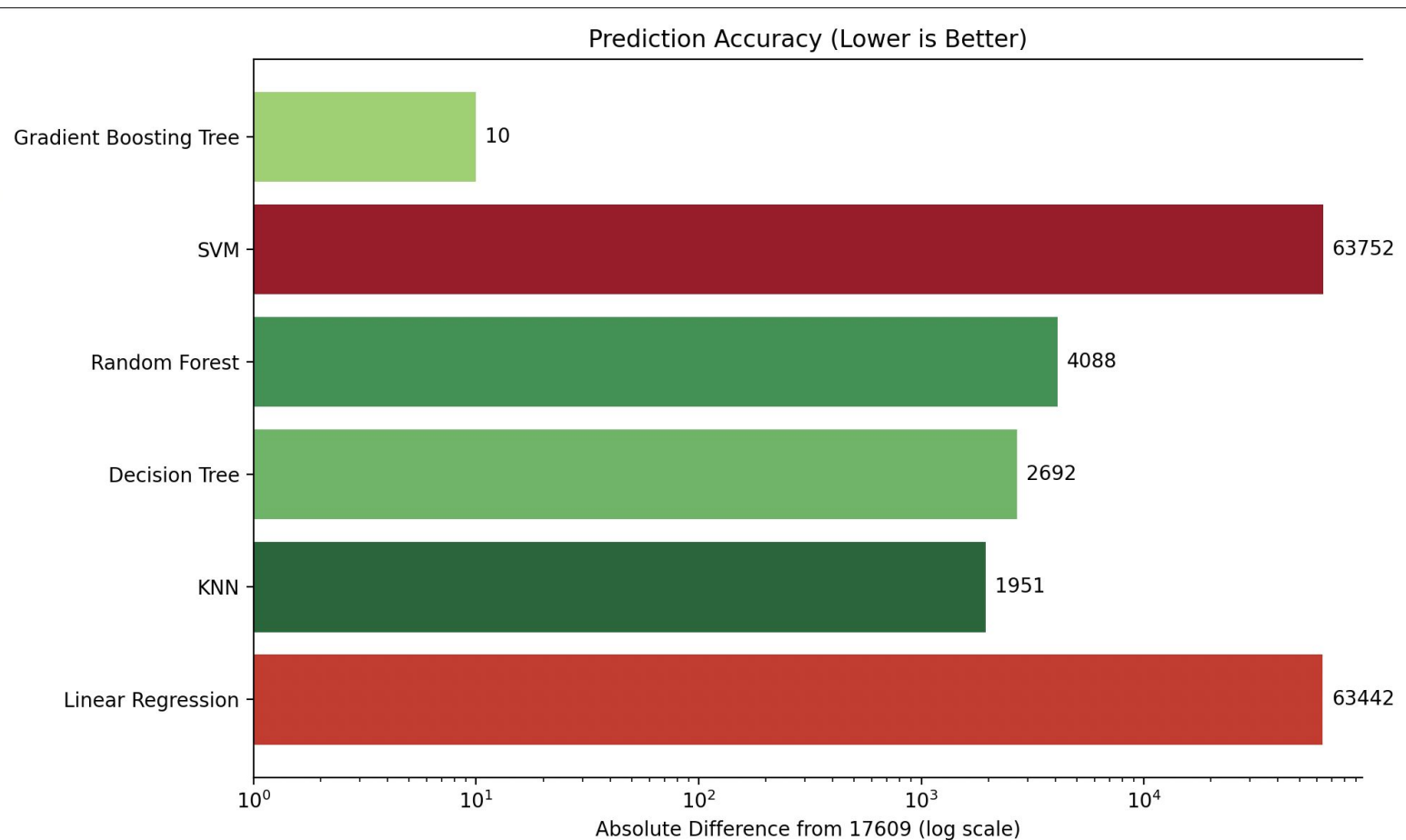
Scenario

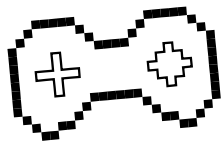


Bob is a farmer from Australia... The year is 2013

He produced: **17609** hg/ha







Thanks!



Hope you enjoyed
Have a nice Tu B'Shvat

By **Eric Gao & Mr.Pham**

