

Farm Yield Prediction Using Machine Learning

-Sharath Yelle



Farm Yield Prediction Using Machine Learning

Step 1: Prototype Selection

ABSTRACT

Farm yield prediction using machine learning is an essential application in modern agriculture that can help farmers optimize resources and improve production efficiency. This project aims to build a predictive model that can accurately forecast farm yields based on various factors such as crop type, farm area, irrigation method, fertilizer usage, pesticide usage, soil type, season, and water usage. Multiple machine learning models, including Linear Regression, Random Forest, XGBoost, CatBoost, and LGBM Regressor, were tested on a dataset, with CatBoost emerging as the best-performing model. The predictive accuracy of the model has the potential to aid farmers in making better decisions regarding resource allocation and crop management, ultimately enhancing productivity and sustainability in agriculture.

1.PROBLEM STATEMENT

Farm yield prediction is a critical challenge for agricultural decision-making. Accurate prediction of crop yield is essential for effective resource management, forecasting food supply, and minimizing risks related to weather conditions, pests, and market fluctuations. Traditional farming methods rely on historical data and expert knowledge, but these can be insufficient and prone to inaccuracies. The goal of this project is to develop a machine learning model that can predict farm yields based on multiple variables, offering a more reliable and data-driven approach to crop yield forecasting.

2.MARKET/CUSTOMER/BUSINESS NEED ASSESSMENT

The agricultural sector is under constant pressure to increase productivity while reducing costs and minimizing environmental impact. Farm yield prediction can significantly benefit various stakeholders:

- **Farmers:** By providing accurate yield predictions, farmers can optimize their resource usage, reduce waste, and plan for harvest, storage, and market sales effectively.
- **Agricultural Businesses:** Companies involved in fertilizer, pesticide, and irrigation system sales can tailor their products to meet specific needs based on the predicted farm yield.

- **Government and NGOs:** Organizations focusing on food security and sustainability can use yield predictions to plan for resource allocation and disaster management, ensuring adequate food supply.
- **Investors and Financial Institutions:** Banks and investors can assess farm risk and make informed decisions about lending and investing in agriculture-based projects.

3.TARGET SPECIFICATIONS AND CHARACTERIZATION

The model aims to predict farm yield with high accuracy, based on the following target specifications:

- **Accuracy:** The prediction model should be able to provide reliable yield estimates with a minimal error margin (e.g., RMSE or MAE).
- **Input Features:** The model will use the following input features: Crop Type, Farm Area, Irrigation Type, Fertilizer Used, Pesticide Used, Soil Type, Season, and Water Usage.
- **Output:** The target variable is **Yield (tons)**, which the model will predict.
- **Data Quality:** The input data should be accurate, up-to-date, and complete, ensuring the model's reliability. Missing or erroneous data may be handled through imputation or other preprocessing techniques.
- **Scalability:** The model should be scalable to larger datasets and adaptable to different crops and regions.

4.EXTERNAL SEARCH(INFORMATION AND DATA ANALYSIS)

These are some of the sources I visited for more information and need for best crop yielding pattern analysis of customers.

- [Crop Yield Increase With Precision Technologies](#)
- [Improvement in Crop Yields](#)
- [Guide to Crop Yield Prediction methods](#)

I am going to use this [Dataset](#) for my code implementation for this report

Dataset Description: Agriculture and Farming Insights

This dataset provides a comprehensive look into the agricultural and farming sector, capturing key variables and trends critical for analysis and decision-making in the

industry. It includes a wide array of data points such as crop types, yield rates, land usage, and farming methods. The dataset is particularly valuable for researchers, policymakers, and agribusiness professionals looking to optimize agricultural output, understand sustainability efforts, or track the economic impacts of farming practices.

5.BENCHMARKING

Benchmarking compares the performance of your predictive models against established baselines or competing algorithms. This process helps to evaluate the effectiveness of your chosen approach.

- **Baseline Model:** A simple model such as **Linear Regression** serves as a baseline for comparison. If it performs well, this indicates that simpler models might suffice for this problem.
- **Random Forest and XGBoost:** Both are commonly used in structured datasets and can model complex relationships between features, providing a strong comparison against the baseline.
- **CatBoost:** Since it outperforms the other models in your experiments, it will serve as the benchmark for optimal performance. You can compare the metrics such as RMSE, MAE, and R^2 for each model to understand their relative strengths.
- **Feature Importance:** Benchmark the models based on their ability to identify the most influential features in predicting farm yield, helping to validate the relevance of the chosen input features.
- **Cross-Validation:** Ensure each model is tested through cross-validation, providing a more generalizable comparison of performance across different subsets of the data.

First import the basic libraries for data preprocessing:

1. Imports and Setup

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display

from sklearn.preprocessing import OneHotEncoder, LabelEncoder

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import RobustScaler

from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from lightgbm import LGBMRegressor
from catboost import CatBoostRegressor
from sklearn.metrics import mean_squared_error, r2_score

import warnings
warnings.filterwarnings("ignore")
```

Let's now see more info on our dataset:

2. Load and Explore Dataset

```
[3] # Load the dataset
df = pd.read_csv('/content/agriculture_dataset.csv')
```

df.head()

	Farm_ID	Crop_Type	Farm_Area(acres)	Irrigation_Type	Fertilizer_Used(tons)	Pesticide_Used(kg)	Yield(tons)	Soil_Type	Season	Water_Usage(cubic meters)
0	F001	Cotton	329.40	Sprinkler	8.14	2.21	14.44	Loamy	Kharif	76648.20
1	F002	Carrot	18.67	Manual	4.77	4.36	42.91	Peaty	Kharif	68725.54
2	F003	Sugarcane	306.03	Flood	2.91	0.56	33.44	Silty	Kharif	75538.56
3	F004	Tomato	380.21	Rain-fed	3.32	4.35	34.08	Silty	Zaid	45401.23
4	F005	Tomato	135.56	Sprinkler	8.33	4.48	43.28	Clay	Zaid	93718.69

```
[ ] df.info()
```

```
>>> <class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 10 columns):
#   Column                      Non-Null Count  Dtype
---  ---                      ---
0   Farm_ID                    50 non-null     object
1   Crop_Type                  50 non-null     object
2   Farm_Area(acres)           50 non-null     float64
3   Irrigation_Type            50 non-null     object
4   Fertilizer_Used(tons)       50 non-null     float64
5   Pesticide_Used(kg)         50 non-null     float64
6   Yield(tons)                50 non-null     float64
7   Soil_Type                  50 non-null     object
8   Season                     50 non-null     object
9   Water_Usage(cubic meters)  50 non-null     float64
dtypes: float64(5), object(5)
memory usage: 4.0+ KB
```

```
df.describe()
```

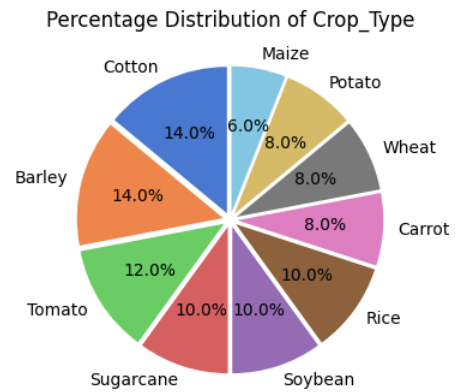
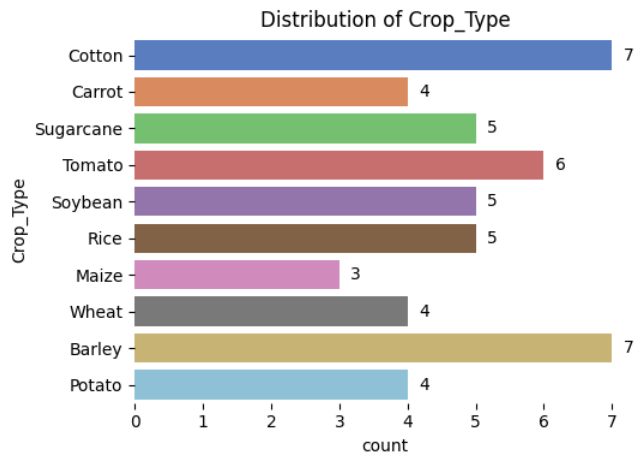
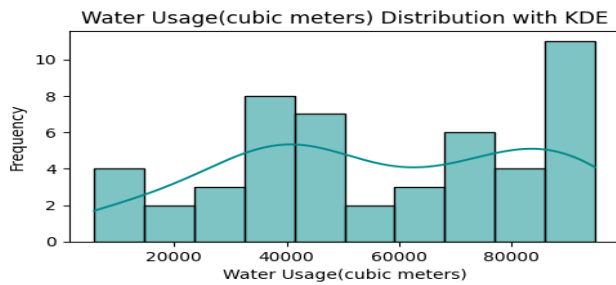
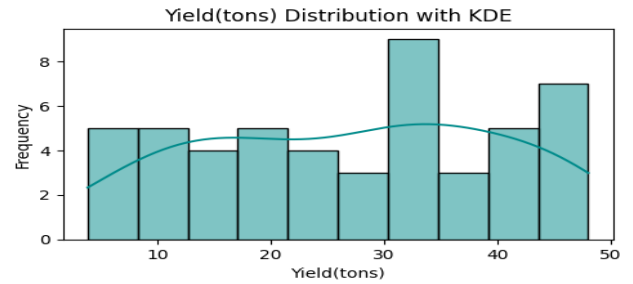
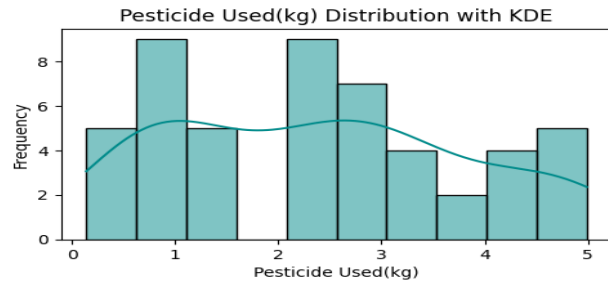
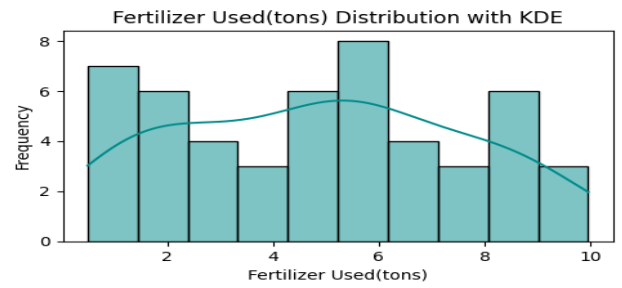
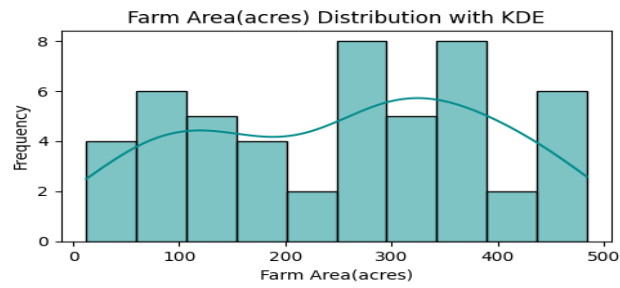
```
>>>
```

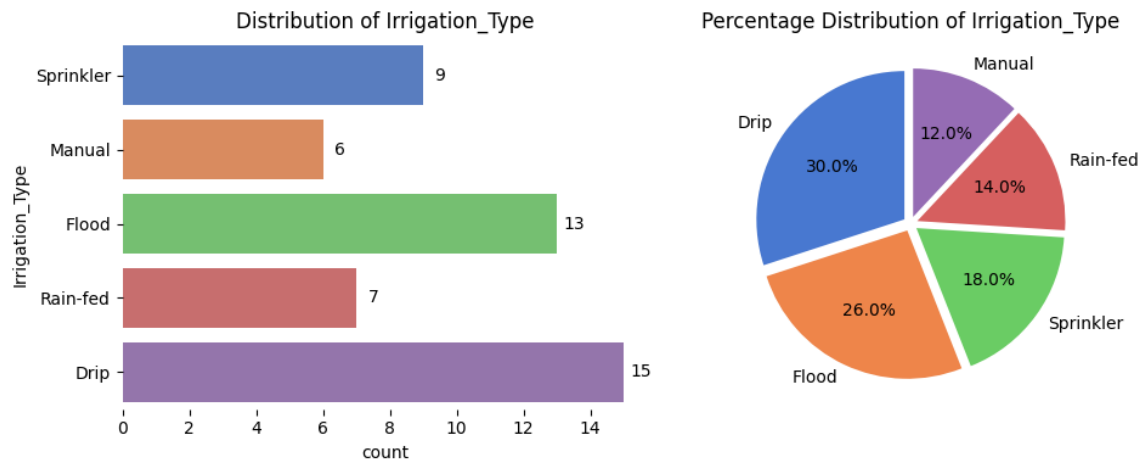
	Farm_Area(acres)	Fertilizer_Used(tons)	Pesticide_Used(kg)	Yield(tons)	Water_Usage(cubic meters)
count	50.000000	50.000000	50.000000	50.000000	50.000000
mean	254.963800	4.905400	2.398000	27.059200	56724.295600
std	139.417782	2.732689	1.438613	13.345789	27264.992053
min	12.500000	0.500000	0.140000	3.860000	5869.750000
25%	135.710000	2.437500	0.972500	16.190000	37818.152500
50%	281.980000	5.045000	2.330000	28.970000	54097.075000
75%	368.107500	6.885000	3.417500	37.860000	82240.032500
max	483.880000	9.960000	4.990000	48.020000	94754.730000

3. Unique Values

```
[6] # number of unique values in each column
print(df.nunique())
```

```
>>> Farm_ID                    50
Crop_Type                    10
Farm_Area(acres)             50
Irrigation_Type               5
Fertilizer_Used(tons)         49
Pesticide_Used(kg)           46
Yield(tons)                   50
Soil_Type                     5
Season                        3
Water_Usage(cubic meters)    50
dtype: int64
```





6.APPLICABLE PATENTS

- Agriculture compositions and applications utilizing mineral compounds
- Novel agricultural composition
- New nutritional and enriching composition for crops

7.APPLICABLE REGULATIONS(GOVERNMENT AND ENVIRONMENTAL)

A. Government Regulations:

- **Data Privacy:** Compliance with GDPR, CCPA for handling personal data.
- **Agricultural Policies:** Adhere to USDA regulations and FAO guidelines for farming practices.
- **Export/Import:** Ensure compliance with export regulations for agricultural technologies.
- **Local/Regional Regulations:** Follow regional rules on pesticide use, water conservation, and crop management.

B. Environmental Regulations:

- **EPA Regulations:** Ensure sustainable water and pesticide use, and minimize air pollution from farming.
- **Sustainable Agriculture:** Align with SAI, organic certification standards, and fair trade practices.

- **Climate Change:** Consider the Paris Agreement, reducing carbon footprints, and integrating climate-friendly practices.
- **Biodiversity Laws:** Follow conservation laws and ensure farming practices support biodiversity.

C. Ethical Considerations:

- **Equity:** Ensure access to predictive technology for all farmers, large or small.
- **Transparency:** Maintain transparency in model predictions and avoid biases in decision-making.
- **Fairness:** Prevent biases in favoring certain farming types, regions, or crops.

8.APPLICABLE CONSTRAINTS

A. Data-Related Constraints:

- **Data Availability:** Limited access to diverse, comprehensive datasets.
- **Data Quality:** Missing, incomplete, or noisy data.
- **Feature Representation:** Inconsistent measurement of features (e.g., soil, water usage).

B. Model Constraints:

- **Computational Resources:** High resource requirements for training complex models.
- **Overfitting:** Risk of poor generalization to unseen data.
- **Model Interpretability:** Difficulty explaining predictions in black-box models.

C. Environmental Constraints:

- **Climate Variability:** Unpredictable environmental factors affecting model accuracy.
- **Sustainability:** Balancing yield predictions with sustainable farming practices.

D. Regulatory Constraints:

- **Data Privacy:** Compliance with GDPR, CCPA if personal data is used.

- **Agricultural Regulations:** Adherence to rules on pesticide, water use, and crop management.

E. Economic Constraints:

- **Cost of Implementation:** High infrastructure and training costs for deployment.
- **Market Dynamics:** Unpredictable market conditions affecting yield predictions.

F. Ethical and Social Constraints:

- **Bias in Predictions:** Ensuring fairness for all regions and farm sizes.
- **Accessibility:** Ensuring model use is available to small-scale farmers.

9. BUSINESS OPPORTUNITY

A. Precision Agriculture Solutions:

- **Optimized Resource Management:** AI-driven insights for efficient water, fertilizer, and pesticide use.
- **Sustainability Consulting:** Help businesses align with sustainable farming practices.
- **Yield Forecasting Service:** Subscription service for real-time crop yield predictions.

B. Data-Driven Decision Support Tools:

- **Farm Management Platforms:** Integrate yield predictions with farm management systems.
- **Market Prediction & Pricing:** Combine yield predictions with market trends for crop planning.
- **Mobile Apps for Farmers:** Provide small farmers with easy access to predictions and recommendations.

C. Partnership with Agricultural Companies:

- **Agri-Tech Startups:** Integrate models into precision farming platforms.
- **Farming Equipment Manufacturers:** Partner to enhance farm equipment with predictive models.

D. Government and NGO Collaborations:

- **Government Programs:** Support agricultural policy-making with yield predictions.
- **NGOs & Rural Development:** Help small farmers increase productivity through affordable tech.

10. CONCEPT GENERATION

- **Farm Management App:** An app providing yield predictions and farm management tips.
- **Precision Irrigation System:** AI-based irrigation scheduling to optimize water use based on yield forecasts.
- **Sustainability Dashboard:** Track sustainability metrics alongside yield predictions.
- **Market Intelligence Platform:** Combine yield predictions with market trends for smarter crop planning.
- **Smart Sensor Network:** IoT sensors for real-time soil and crop data to improve predictions.
- **Agricultural Data Marketplace:** Platform for farmers to sell/share their data for better predictions.

11. CONCEPT DEVELOPMENT

steps to develop a concept for your **Farm Yield Prediction** project:

A. Define Core Objectives:

- **Objective:** Develop a solution that predicts farm yields with high accuracy using machine learning models like CatBoost and other tools, improving decision-making for farmers.

B. Target Audience:

- **Farmers:** Small to large-scale farmers looking for insights to optimize crop yield and resource usage.

- **Agricultural Organizations:** NGOs, government bodies, and agricultural businesses aiming to support farmers with data-driven solutions.
- **Agri-Tech Companies:** Tech firms interested in integrating prediction models into farm management systems.

C. Key Features:

- **Yield Prediction:** Provide accurate yield predictions based on farm data (e.g., crop type, soil quality, water usage).
- **Resource Optimization:** Suggest optimal water, fertilizer, and pesticide use to maximize yield and minimize cost.
- **Market Forecasting:** Include market trends and pricing predictions to guide farmers on what crops to grow and when to harvest.
- **Sustainability Tracking:** Measure and recommend eco-friendly practices (e.g., water conservation, sustainable fertilization).
- **Mobile Interface:** User-friendly mobile app for easy access to predictions and recommendations.

D. Feasibility Analysis:

- **Technical Feasibility:** Use machine learning models (like CatBoost) and real-time weather data to generate predictions. Ensure data availability for training models.
- **Economic Feasibility:** Estimate initial costs for developing the system, including data acquisition, software development, and training. Evaluate pricing models for farmers.
- **Operational Feasibility:** Design an easy-to-use interface that farmers with limited technical knowledge can operate. Ensure scalability for different farm sizes and regions.

E. Development Roadmap:

1. **Phase 1 - Data Collection & Preprocessing:** Gather data on crop yield, weather, soil quality, and other relevant features.

2. **Phase 2 - Model Training & Testing:** Implement and evaluate models like CatBoost, XGBoost, and others to find the best-performing model for yield prediction.
3. **Phase 3 - Platform Development:** Build a web or mobile platform to integrate the model, allowing farmers to input data and receive predictions and recommendations.
4. **Phase 4 - Market Launch:** Introduce the solution to the target audience through collaborations with agricultural organizations, NGOs, and tech companies.
5. **Phase 5 - Continuous Improvement:** Collect feedback, refine the model, and enhance the platform based on real-world use.

F. Risk Assessment:

- **Data Quality:** Inaccurate or incomplete data could impact the prediction accuracy. Regular updates and data validation mechanisms are essential.
- **Model Overfitting:** Overfitting on training data can reduce the model's generalization. Regularization techniques and cross-validation should be used.
- **Adoption Challenges:** Farmers may be hesitant to adopt new technology. Training programs and user-friendly designs will help overcome resistance.

G. Future Enhancements:

- **Integration with IoT Devices:** Incorporate IoT sensors for real-time farm monitoring to improve prediction accuracy.
- **AI-Powered Decision Support:** Introduce more advanced decision-making tools, like automated irrigation systems based on yield forecasts.
- **Global Expansion:** Adapt the solution to different geographical regions with varying crop types, climates, and farming practices.

12.FINAL PRODUCT PROTOTYPE/ PRODUCT DETAILS

- **Product: Farm Yield Prediction System**
- **Core Features:**

- **Yield Prediction:** Accurate predictions based on crop type, soil, water usage, etc., using machine learning (CatBoost).
- **Resource Optimization:** Suggestions for optimal irrigation, fertilizer, and pesticide use.
- **Market Forecasting:** Predicts crop prices and harvest timing for better financial planning.
- **Sustainability Tracker:** Measures environmental impact (water, pesticide use) and recommends eco-friendly practices.
- **Mobile & Web Platforms:** User-friendly interfaces for farmers to access predictions, resource tips, and alerts.
- **Technology:**
 - **ML Models:** CatBoost, XGBoost, etc.
 - **Web:** React for frontend, Flask for backend.
 - **Mobile:** React Native/Flutter for cross-platform apps.
 - **Cloud:** AWS/Google Cloud for hosting and processing.
- **User Flow:**
 1. Register and input farm data.
 2. Receive predictions and recommendations.
 3. Get alerts for critical actions (irrigation, harvesting).
- **Future Features:**
 - IoT integration for real-time data.
 - AI-powered decision support.
 - Global market data and multilingual support.
- **Pricing:** Subscription-based model, with freemium options for basic features.

STEP 2: PROTOTYPE DEVELOPMENT

GitHub Link : [Github Repo Link](#)

STEP 3: BUSINESS MODELING

- **Value Proposition:**
 - **Farmers:** Accurate yield predictions, resource optimization, and market forecasting to maximize profits.
 - **Agri-Tech Companies & NGOs:** Data-driven insights for improved farming practices and sustainability.
- **Revenue Model:**
 - **Subscription-Based:** Monthly/annual fees for different farm sizes and premium features.
 - **Freemium:** Basic access free, premium features behind a paywall.
 - **Data Licensing:** Sell anonymized data for agricultural research.
- **Key Resources:**
 - Machine learning models (CatBoost, XGBoost).
 - Cloud infrastructure and web/mobile app development.
 - Data scientists, software developers, and support staff.
- **Key Activities:**
 - Train models, update platform, and provide customer support.
 - Conduct market research and customer engagement.
- **Cost Structure:**
 - Development, operational, and sales costs.
 - Cloud hosting, data storage, and maintenance.
- **Target Customers:**
 - Small to large-scale farmers, agricultural NGOs, and agri-tech companies.
- **Marketing Strategy:**

- Direct sales, digital marketing, partnerships with NGOs and government bodies.
- **Key Success Factors:**
 - Accurate predictions, user-friendly interface, scalability, and data security.
- **Risks & Mitigation:**
 - **Data Accuracy:** Regular data validation.
 - **Farmer Adoption:** Training and easy-to-use design.

STEP 4: FINANCIAL MODELING

1. To calculate the **total revenue** and **profit** for the **Farm Yield Prediction System**, we'll follow the basic financial equation:

- **Total Revenue (y)** = (Price per Unit) * (Number of Units Sold) - (Cost to Produce/Operate)

Example Assumptions:

- **Price per Unit:** Rs. 500 per month (Subscription fee per farm)
- **Operating Cost:** Rs. 2000 per month (Fixed cost)
- **Number of Units Sold (x):** Number of farms subscribing in a month

Equation:

- **Revenue (y)** = $500 * x - 2000$

Where:

- **x** = Number of farms subscribing in the month
- **y** = Total monthly revenue

Example Calculation:

- If 300 farms subscribe:
 - **Total Revenue** = $500 * 300 - 2000 = \text{Rs. } 1,49,800$

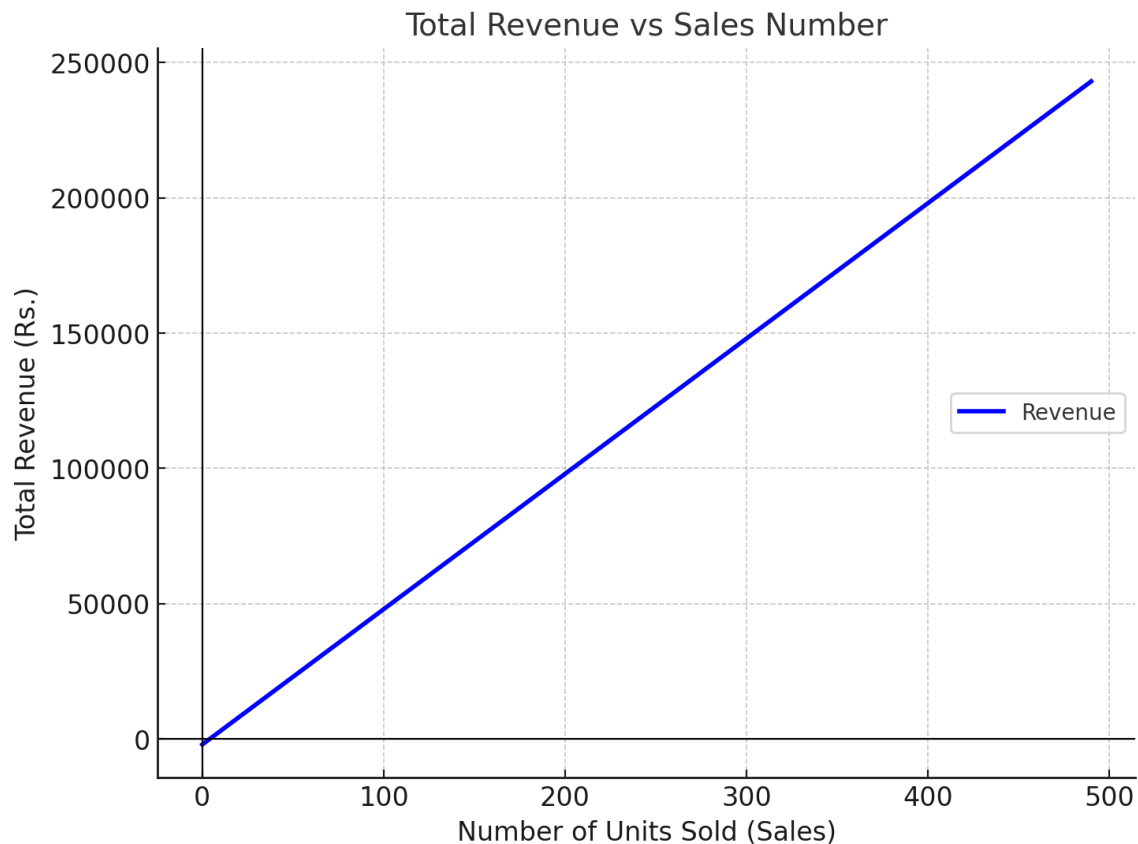
This basic model can be expanded further by including variables like customer acquisition cost, marketing expenses, and growth projections to get a more comprehensive financial outlook.

Let me know if you'd like a more detailed breakdown!

2. You said:

give me a graph

3. ChatGPT said:



Here is the graph representing **Total Revenue** as a function of **Sales Numbers** for the Farm Yield Prediction System. The x-axis shows the number of units (farms subscribing), and the y-axis shows the total revenue (Rs.).

As the number of sales increases, the total revenue rises, indicating that the business generates more revenue with more subscriptions, while considering the fixed operating cost.