CORN YIELD FORECAST

Parsa Taati
Afsoun Abbasi
Hamed Goldoust

FONDAZIONE LINKS
PASSION FOR INNOVATION
Politecnico di Torino
1859

Project Value Preposition

For farmers, internal team members, and academic advisors, who require actionable insights for crop management, reliable datasets for accurate model development, and well-documented progress for academic evaluation,
our project, **Corn Yield Forecast**, is a data-driven decision-support system leveraging predictive models and weather analytics, that provides optimal planting and harvesting recommendations, clean and accessible datasets, and comprehensive project documentation to enhance agricultural efficiency, streamline development processes, and ensure academic rigor.
Unlike existing solutions, our system integrates advanced predictive analytics with user-centric design, ensuring accurate yield forecasting, seamless data accessibility, and actionable insights tailored to the unique needs of farmers, researchers, and development teams, while promoting sustainability and academic excellence.

General Objectives

Develop Accurate Corn Yield Predictions

Optimize Resource Allocation

Incorporate Weather Analytics

Enhance Agricultural Decision-Making

Research Questions

Which features are most relevant for accurately predicting corn yield?

How do environmental variables (e.g., temperature, precipitation, and solar radiation) correlate with corn yield?

Pipeline for Corn Yield Prediction:

1

Row Data

2

Preprocessing

- Data cleaning
- handling missing values

3

Feature Selection

- Analyzing correlations
- selecting the most relevant features

4

Model Training

- Training the Random Forest model
- Logistic Regression

5

Evaluation

- Assessing model performance
- generating yield predictions

Merged dataset features:

- YEAR
- FAO Cicle
- Month
- Frost Days
- Heavy Rain Events
- Average Temperature
- Maximum Temperature
- Minimum Temperature
- Growing Degree Days
- Heat Stress Days
- Cold Stress Days
- Temperature Variability
- Total Precipitation
- Rainy Days
- Drought Days
- Maximum Daily Rainfall
- Rainfall Variability
- Consecutive Dry Days
- Cumulative Solar Radiation
- Average Solar Radiation
- Potential Evapotranspiration
- Average Wind Speed
- Maximum Wind Speed
- Average Relative Humidity
- Humidity Variability

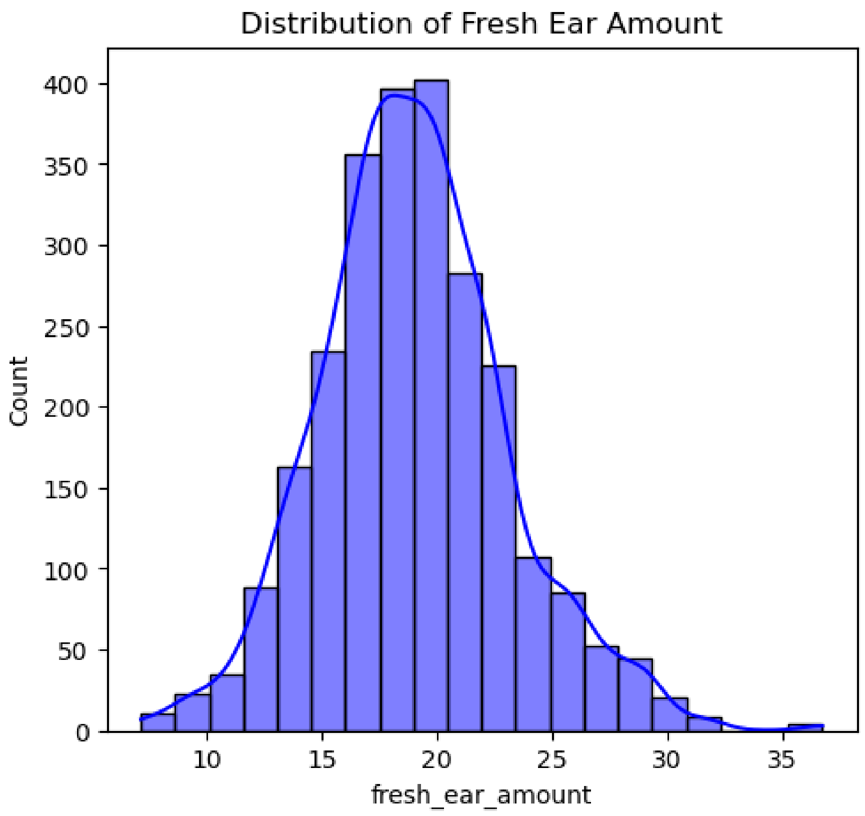
Target values

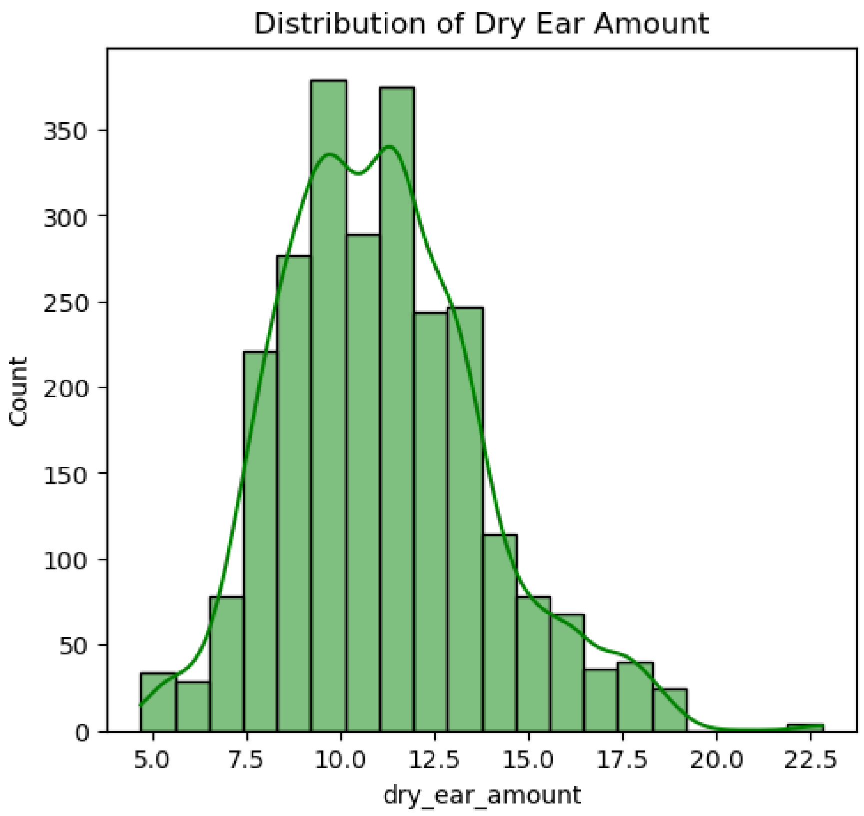
Fresh Ear yield = (Fresh Ear Percentage/100)×Fresh Total

Dry Ear yield =(Dry Ear Percentage/100)×Dry Total

EDA in Preprocessing

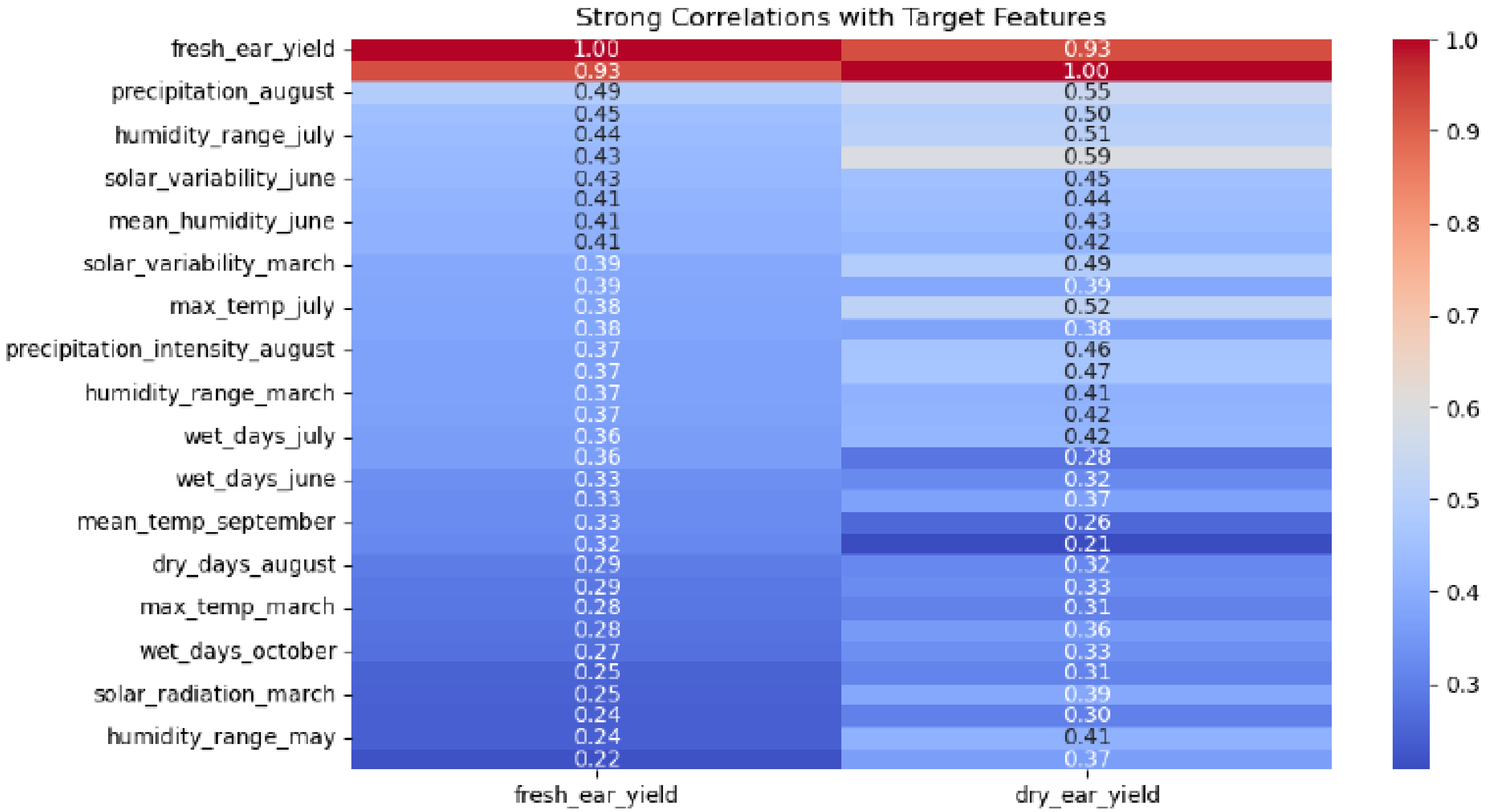
Outlier Detection
Handling Missing value
Normalization





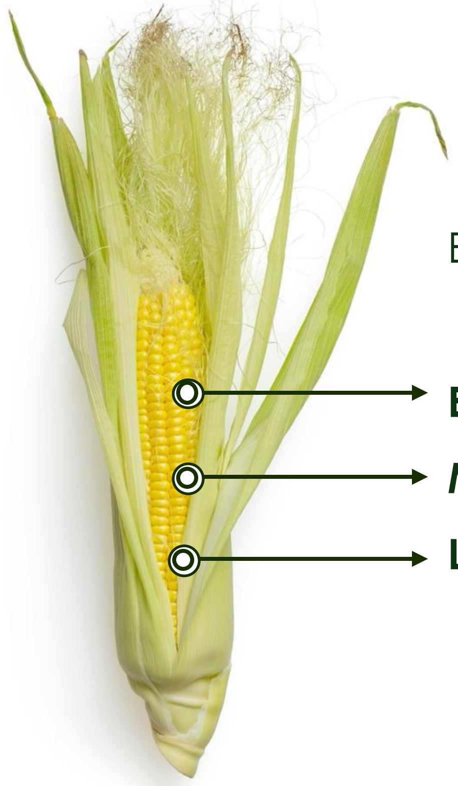
Which are the most relevant features to collect ?

➔ **Correlation Matrix**



Importance of precipitation and time series characteristic:

➔ **We need to aggregate daily information to the month level**



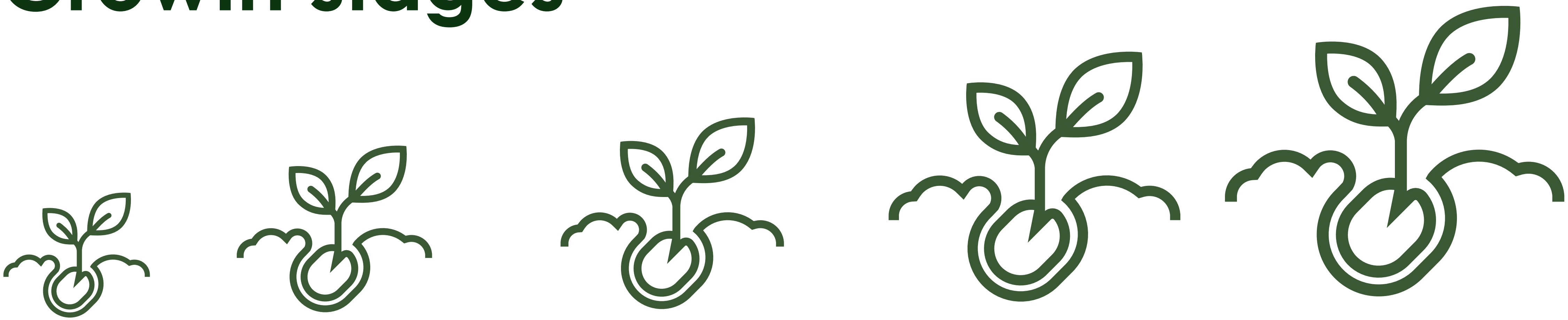
By considering the corn life cycle we reached that

Early Stage: Germination and emergence.

Mid-Stage: Vegetative growth , tasseling , and silking.

Late Stage: Kernel development and physiological maturity.

Growth stages



Planting
March-May

Vegetative Growth
April-June

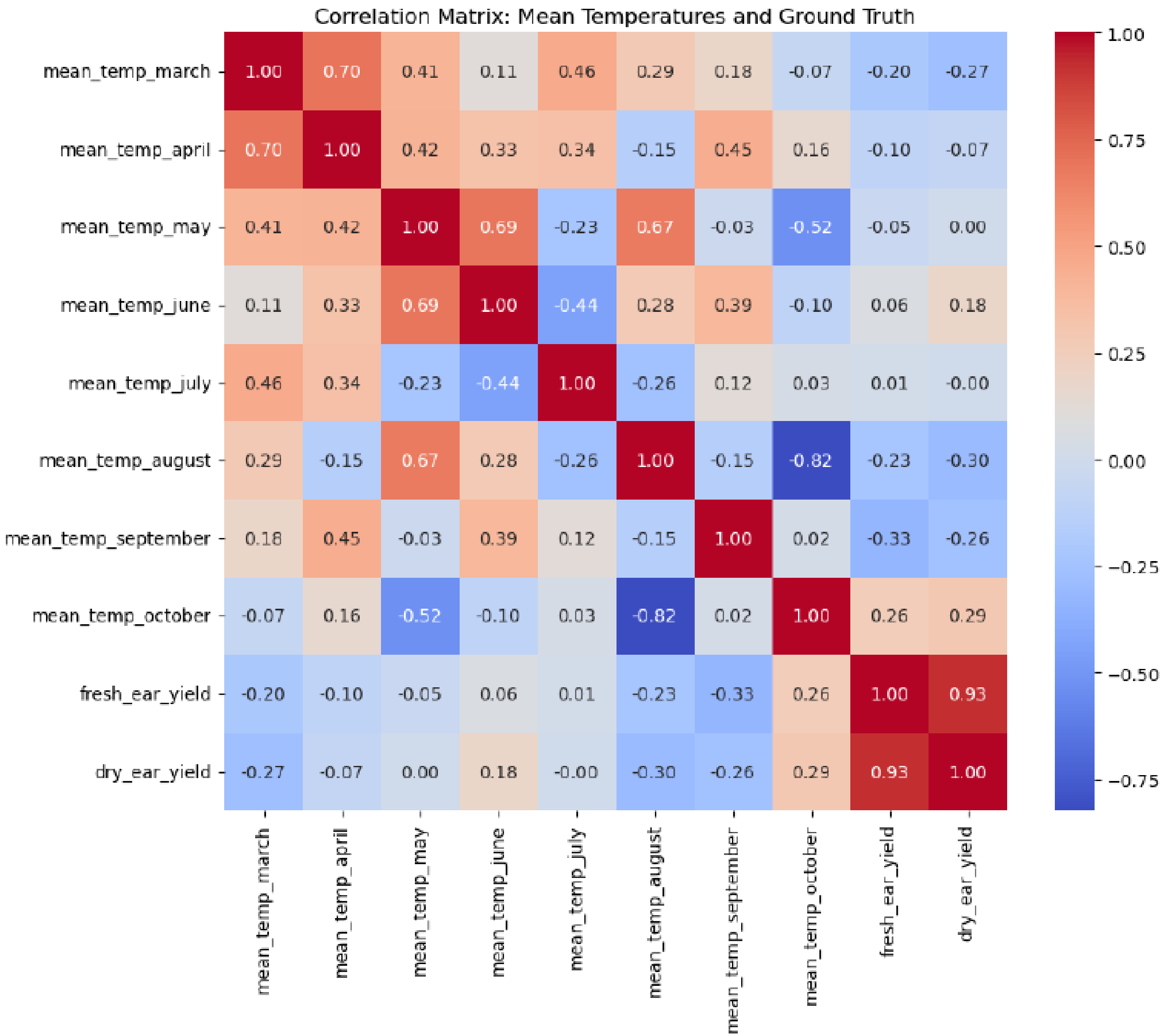
Tasseling and Silking
June-July

Kernel Development
July-August

Maturity and Harvest
September-October

Then we diminish months which does not have impact on corn

Corrolation of important features with lables



Dataset Split

TRAIN SET

2011 2020

2021 2023

TEST SET

Methods:

RandomForest Regressor

XGBoost

Support Vector Machines (SVR)

Gradient Boosting

Regression Models:

Lasso Regression

Ridge Regression

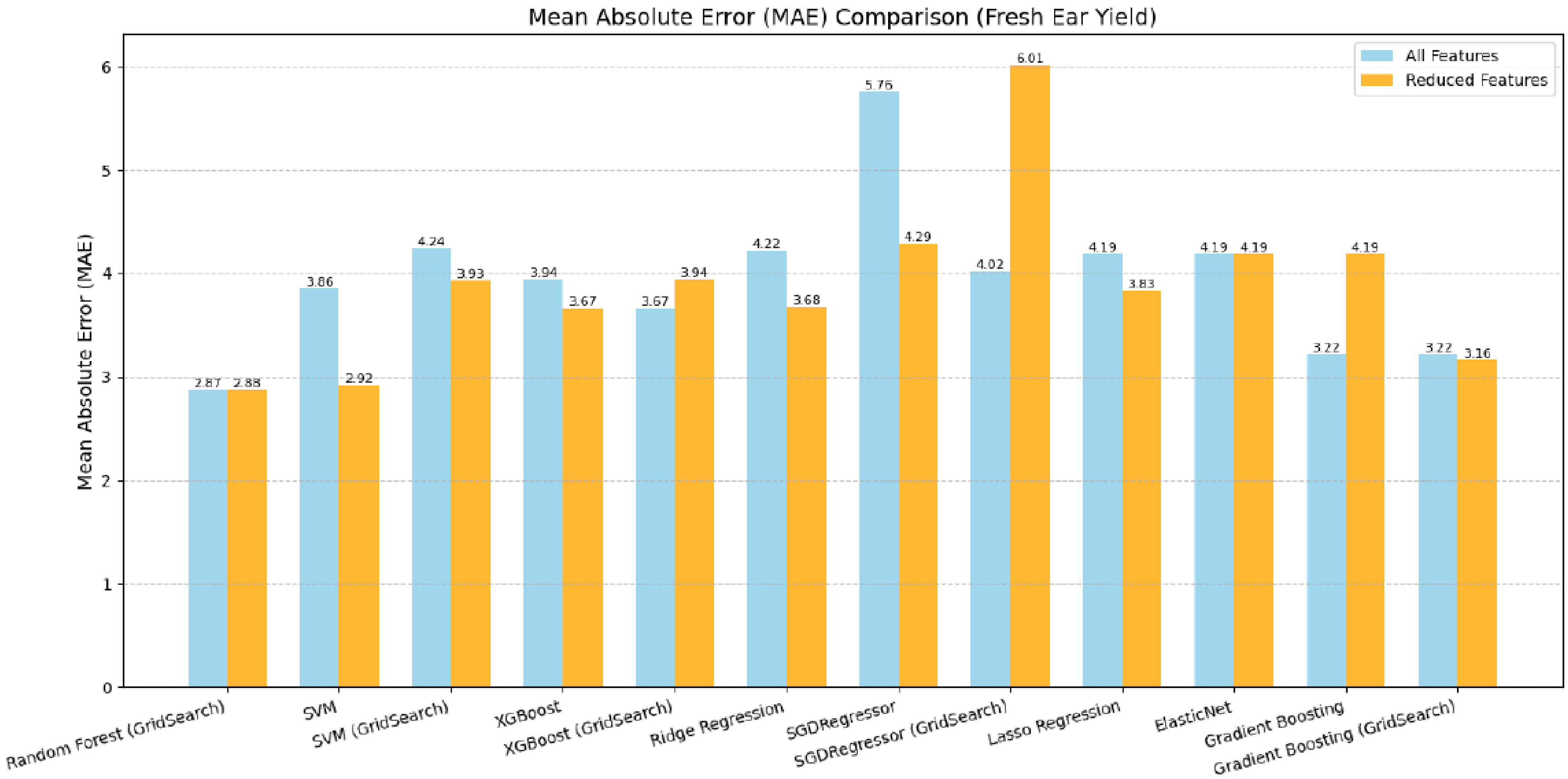
ElasticNet Regression

Evaluation Metrics:

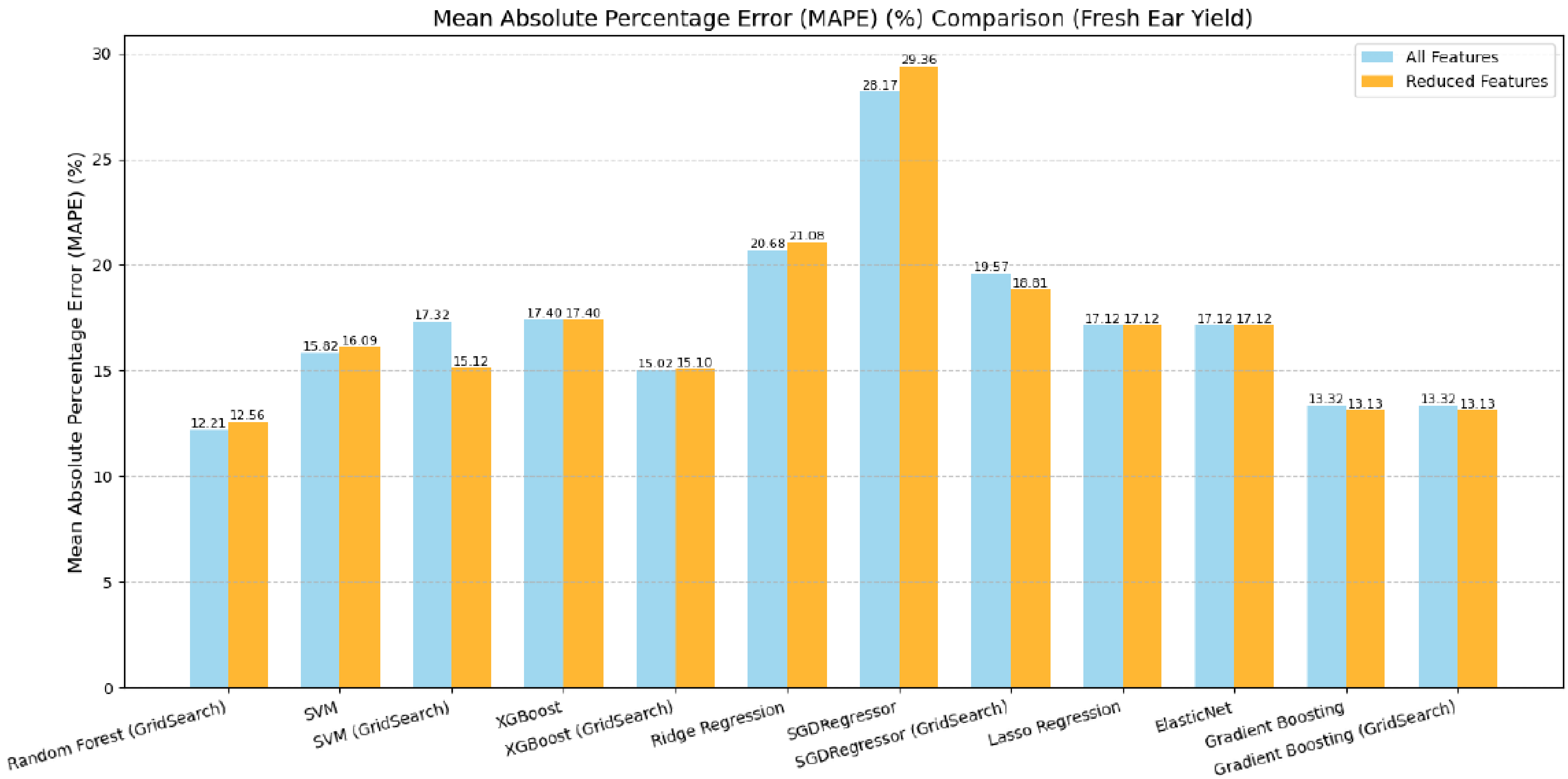
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

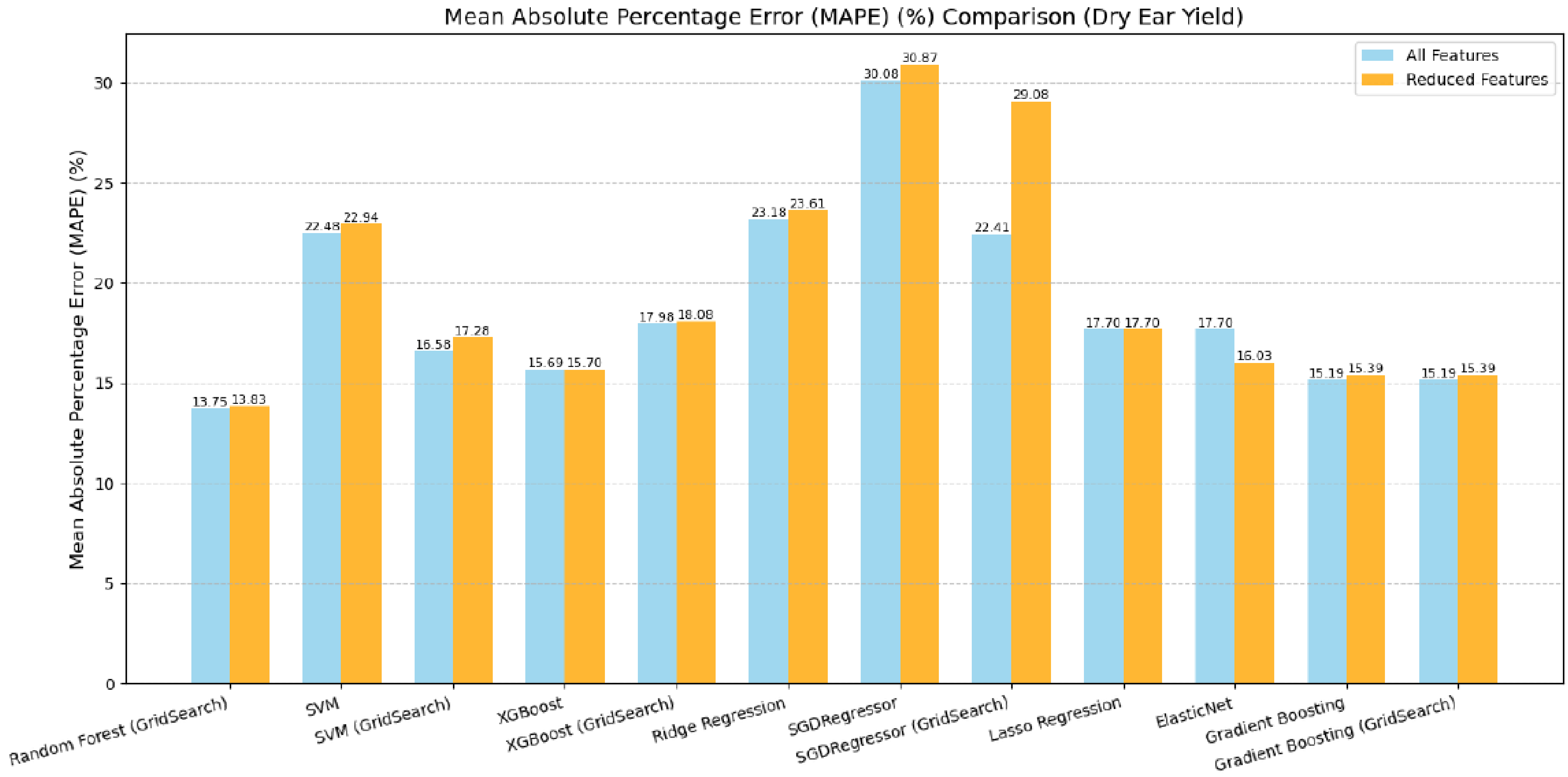
$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

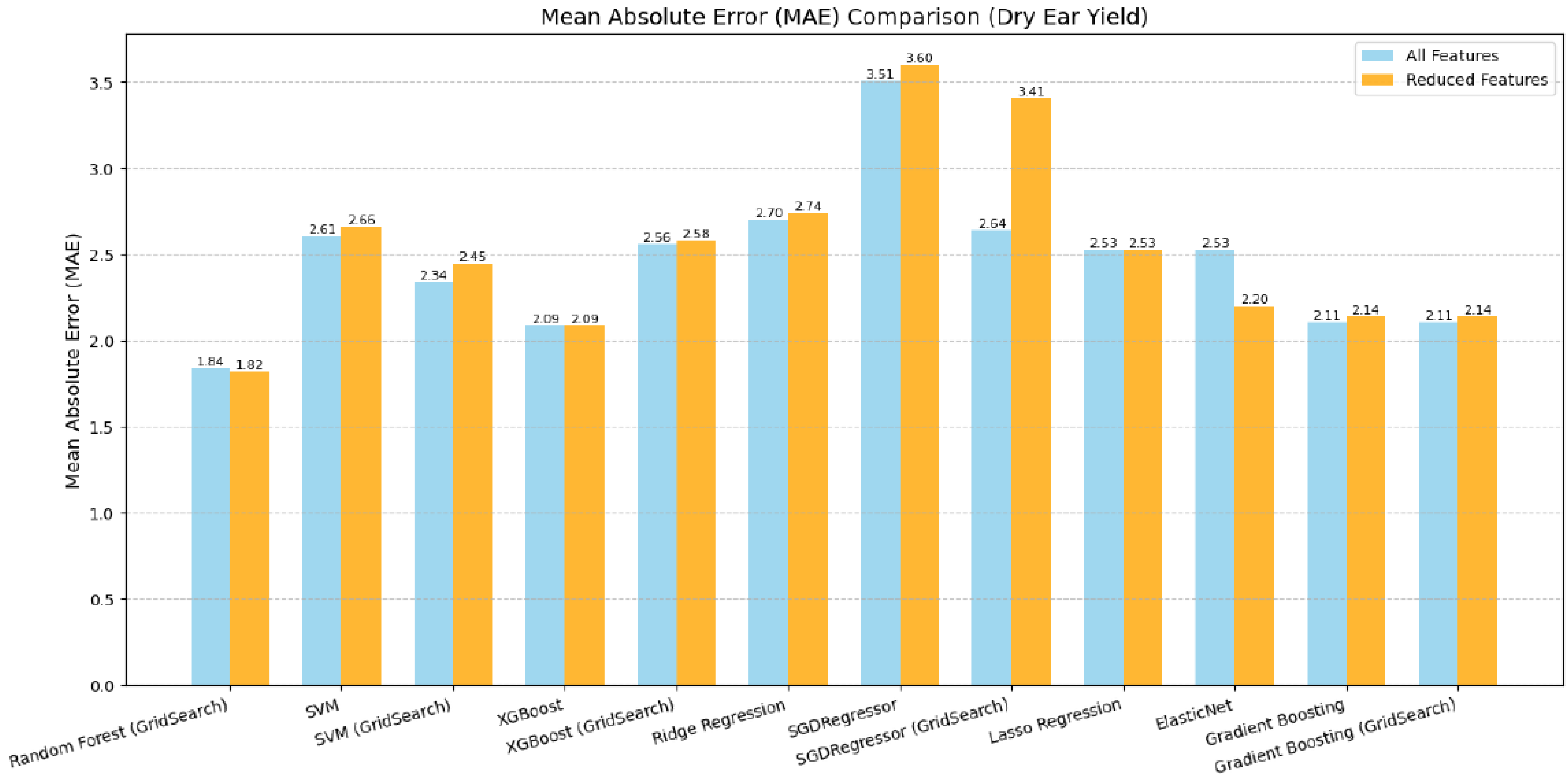
Experiments and Conclusion



Mean Absolute Error (MAE) Comparison (Fresh Ear Yield)
All Features
Reduced Features
Mean Absolute Error (MAE)
0
1
2
3
4
5
6
2.87
2.88
3.86
2.92
4.24
3.93
3.94
3.67
3.67
3.94
4.22
3.68
5.76
4.29
4.02
6.01
4.19
3.83
4.19
4.19
3.22
4.19
3.22
3.16
Random Forest (GridSearch)
SVM
SVM (GridSearch)
XGBoost
XGBoost (GridSearch)
Ridge Regression
SGDRegressor
SGDRegressor (GridSearch)
Lasso Regression
ElasticNet
Gradient Boosting
Gradient Boosting (GridSearch)


Mean Absolute Percentage Error (MAPE) (%) Comparison (Fresh Ear Yield)
Mean Absolute Percentage Error (MAPE) (%)
All Features
Reduced Features
0
5
10
15
20
25
30
12.21
12.56
15.82
16.09
17.32
15.12
17.40
17.40
15.02
15.10
20.68
21.08
28.17
29.36
19.57
18.81
17.12
17.12
17.12
17.12
13.32
13.13
13.32
13.13
Random Forest (GridSearch)
SVM
SVM (GridSearch)
XGBoost
XGBoost (GridSearch)
Ridge Regression
SGDRegressor
SGDRegressor (GridSearch)
Lasso Regression
ElasticNet
Gradient Boosting
Gradient Boosting (GridSearch)


Mean Absolute Percentage Error (MAPE) (%) Comparison (Dry Ear Yield)
All Features
Reduced Features
Mean Absolute Percentage Error (MAPE) (%)
0
5
10
15
20
25
30
13.75
13.83
22.48
22.94
16.58
17.28
15.69
15.70
17.98
18.08
23.18
23.61
30.08
30.87
22.41
29.08
17.70
17.70
17.70
16.03
15.19
15.39
15.19
15.39
Random Forest (GridSearch)
SVM
SVM (GridSearch)
XGBoost
XGBoost (GridSearch)
Ridge Regression
SGDRegressor
SGDRegressor (GridSearch)
Lasso Regression
ElasticNet
Gradient Boosting
Gradient Boosting (GridSearch)


Mean Absolute Error (MAE) Comparison (Dry Ear Yield)
Mean Absolute Error (MAE)
All Features
Reduced Features
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1.84
1.82
2.61
2.66
2.34
2.45
2.09
2.09
2.56
2.58
2.70
2.74
3.51
3.60
2.64
3.41
2.53
2.53
2.53
2.20
2.11
2.14
2.11
2.14
Random Forest (GridSearch)
SVM
SVM (GridSearch)
XGBoost
XGBoost (GridSearch)
Ridge Regression
SGDRegressor
SGDRegressor (GridSearch)
Lasso Regression
ElasticNet
Gradient Boosting
Gradient Boosting (GridSearch)

hyperparameter tuning for a Random Forest model

Hyperparameter	Description	Values Explored
`n_estimators`	Number of trees in the forest	{100, 200, 300}
`max_depth`	Maximum depth of the trees	{10, 20, None}
`min_samples_split`	Minimum samples required to split an internal node	{2, 5, 10}
`min_samples_leaf`	Minimum samples required at a leaf node	{1, 2, 4}
`max_features`	Number of features considered for the best split	{"sqrt", "log2", 0.8}
`bootstrap`	Sampling with replacement (True/False)	{True, False}

Table 3: Hyperparameter tuning values explored for a Random Forest model

Results

Model	MAE (Fresh Ear Yield)	MAPE (Fresh Ear Yield) %	MAE (Dry Ear Yield)	MAPE (Dry Ear Yield) %
RandomForest (GridSearch)	**2.87**	**12.21**	**1.84**	**13.75**
SVM	3.86	15.82	2.61	22.48
SVM (GridSearch)	4.24	17.32	2.34	16.58
XGBoost	3.94	17.40	2.09	15.69
XGBoost (GridSearch)	3.67	15.02	2.56	17.98
Ridge Regression	4.22	20.68	2.70	23.18
SGDRegressor	5.76	28.17	3.51	30.08
SGDRegressor (GridSearch)	4.02	19.57	2.64	22.41
Lasso Regression	4.19	17.12	2.53	17.70
ElasticNet	4.19	17.12	2.53	17.70
Gradient Boosting	3.22	13.32	2.11	15.19

Table 1: Evaluation Metrics for Machine Learning Models Using All Features

Model	MAE (Fresh Ear Yield)	MAPE (Fresh Ear Yield) %	MAE (Dry Ear Yield)	MAPE (Dry Ear Yield) %
RandomForest (GridSearch)	**2.92**	**12.56**	**1.82**	**13.83**
SVM	3.93	16.09	2.66	22.94
SVM (GridSearch)	3.67	15.12	2.45	17.28
XGBoost	3.94	17.40	2.09	15.70
XGBoost (GridSearch)	3.68	15.10	2.58	18.08
Ridge Regression	4.29	21.08	2.74	23.61
SGDRegressor	6.01	29.36	3.60	30.87
SGDRegressor (GridSearch)	3.83	18.81	3.41	29.08
Lasso Regression	4.19	17.12	2.53	17.70
ElasticNet	4.19	17.12	2.20	16.03
Gradient Boosting	3.16	13.13	2.14	15.39

Table 2: Evaluation Metrics for Machine Learning Models Using Reduced Features

Thanks