Application of Machine Learning in predicting Crop yield

Elijah Adebimpe
B1152597

School of Computing, Engineering and Digital Technologies(SCEDT)
Teesside University, England, United Kingdom.

***Abstract*— Crop yield forecasting has grown increasingly important in ensuring that the world's food supply is met. The accuracy of five machine learning methods in forecasting yield of ten crops in 101 nations of the globe was evaluated in this work (using R2 and RMSE score). Random forest achieved the best R2 score (0.97), as well as the lowest RMSE making it the best model for large-scale crop yield prediction, according to the research.**

I. INTRODUCTION

According to a study carried out by the UN, global population is projected to grow beyond the productive capacity of the agricultural sector. It is projected that the world populace would hit 9.1 billion in 2050, which is a significant increase from what it is today [5].

This would mean that demand for food would naturally outgrow the supply, and feeding this massive populace based on the current declining productive capacity of the agricultural sector relative to the increase in population becomes a problem that needs to be solved. Agriculture would also have to compete for natural resources such as land and water, with the rapidly growing population settling down. As a result, it becomes a priority to produce more food with less of these natural resources and optimize the production process leveraging on technology.

Forecasting the yield of crops has become an essential field of study that ensures food security globally. Crop yield is a measure of the quantity of agricultural production harvested per unit of land area [1].

Rainfall, temperature, and many other global climate change conditions can affect the yield of these crops. These varying conditions make the accurate prediction of crop yield very challenging when done manually or traditionally [2].

Using Machine Learning Algorithms and relevant crop information contained in data makes accurate crop yield prediction more attainable. Machine learning is a effective approach to predicting yields based on a variety of inputs and can extract information from large datasets by recognizing patterns and correlations.

II. RELEVANT WORKS

Thomas et al. conducted a systematic literature assessment of 50 machine learning articles relevant to crop yield prediction in 2020. The study's findings revealed that the most employed features were temperature, rainfall, and soil type, all of which were linked to soil, sunlight, and humidity data. The study's findings are summarised in the table below [4].

Most used machine learning algorithms	# of times used
Neural Networks	27
Linear Regression	14
Random Forest	12
Support Vector Machine	10
Gradient Boosting Tree	4

Key	Evaluation parameter	# of times used
RMSE	Root mean square error	29
R^2	R-squared	19
MAE	Mean absolute error	8
MSE	Mean square error	5
MAPE	Mean absolute percentage error	3
RSAE	Reduced simple average ensemble	3
LCCC	Lin's concordance correlation coefficient	1
MFE	Multi factored evaluation	1
SAE	Simple average ensemble	1
rcv	Reference change values	1
MCC	Matthew's correlation coefficient	1

Regression tree, random forest, support vector regression, and artificial neural networks were among the machine learning methods employed before by researchers. Input variables utilised include average rainfall, pesticide treatment, and other factors. Although the linear machine learning approach has been widely utilised for crop yield prediction, it has been criticised for its poor accuracy. For example, Dixon et al in 1994 and Sudduth et al in 1996 observed accuracies ranging from poor to moderate when using linear techniques [8].

Fukuda et al. employed random forest to estimate mango fruit yields in response to water availability under various irrigation regimes and found that random forest was a better predictor for mango yields with an emphasis on water management. In 2004, Jiang et al. used artificial neural networks and multiple linear regression to anticipate winter wheat yields using remotely sensed and meteorological data, finding that the artificial neural network model beat the multiple linear regression model. [2].

Joeng et al. undertook research to estimate agricultural yields for wheat, maize, and potatoes. They employed two machine learning algorithms, Random Forest and Multiple Linear Regression, for comparison, with Random Forest being the stronger approach for predicting crop yield. As a

comparison's baseline, the Root Mean Square Error was used (RMSE). Climate, soil, photoperiod, water, and fertiliser data were among the environmental variables examined [7].

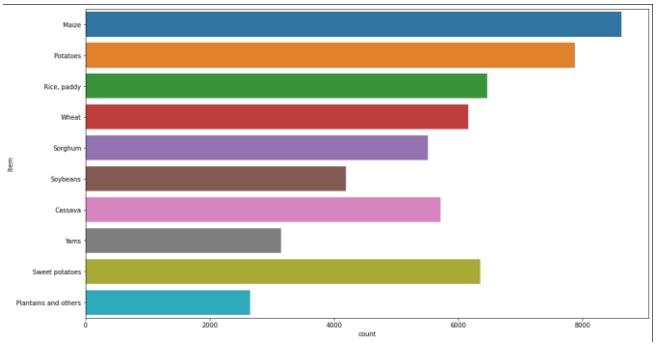
III. ABOUT THE DATA

Separate datasets containing information relating to pesticides, yield, rainfall, and average temperature were downloaded from Kaggle and are also publicly obtainable from the FAO and World Bank Data website. These data sets were merged based on Region and Year and final dataset contained information pertaining to 10 commonly grown crops in 101 regions of the world. The final dataset contained the following columns:

- Area.
- Item
- Year
- Crop yield
- Average rain yearly
- Pesticides
- Average temperature

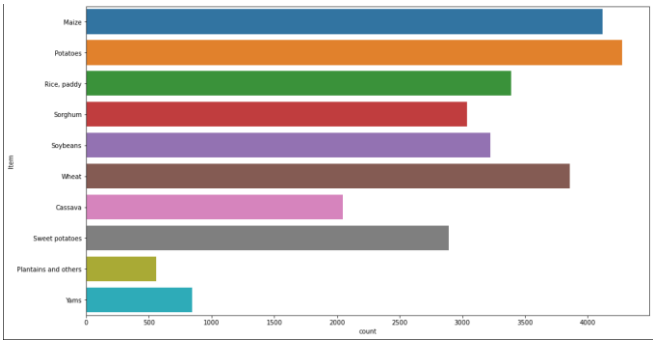
IV. EXPLORATORY DATA ANALYSIS AND PREPARATION

An observation from the yield.csv dataset that contained information as regards the 10 commonly grown crops is shown below:



The visual above show that maize and potatoes were the prominently grown crops in the 212 regions of the world covered in the yield dataset.

However, upon merging the yield dataset with average temperature, rainfall and pesticides use, there was a change in the distribution of the crops covered as shown below;



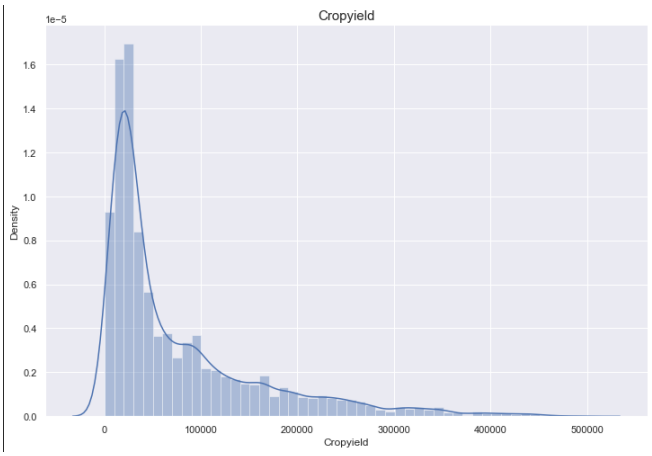
There are more potatoes than maize in the merged dataset. This can be attributed to the reduction in regions covered within the initial yield data and final merged dataset.

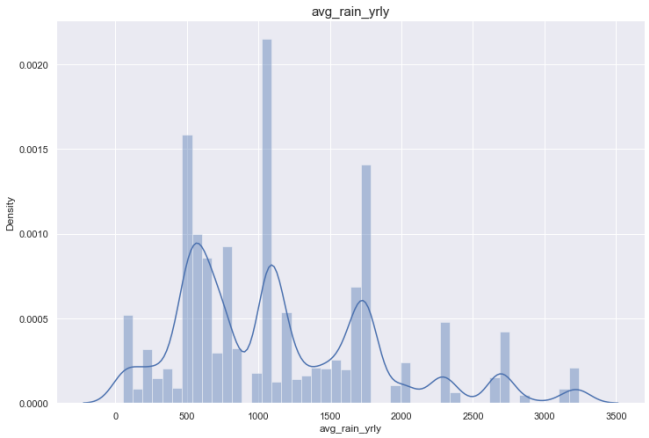
The number of regions covered missing values and number of rows and column in each separate csv is shown below;

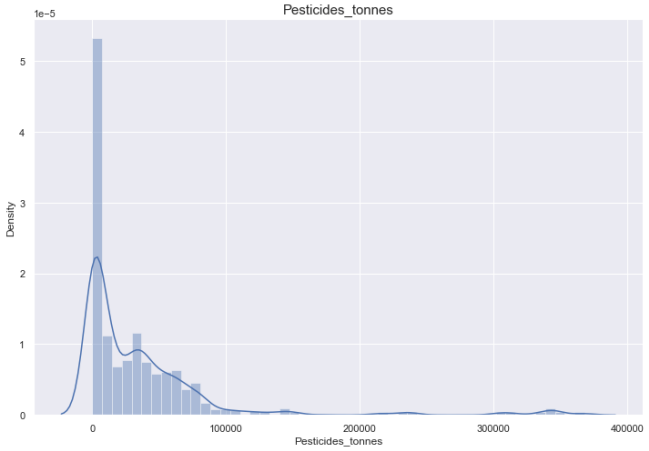
	Regions covered in each dataset		
Dataset	***Number of Rows and Columns***	***Regions covered***	***Number of missing values***
Yield	56,717 and 12	212	8
Rainfall	6,727 and 3	217	780
Pesticide	4349 and 7	168	2
Average temperature	71,311 and 3	137	2547

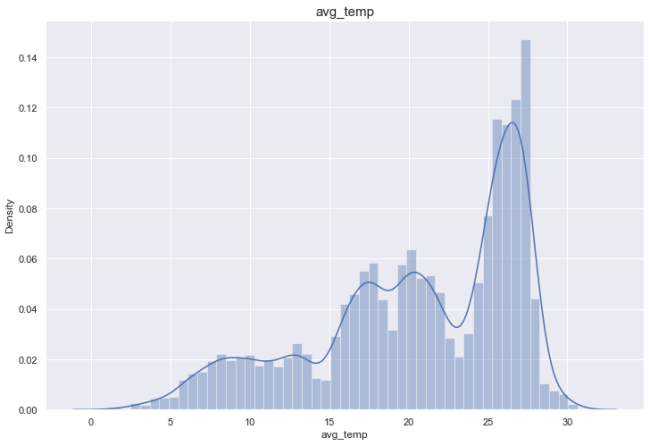
Common to all csv files were the Year and Area (Region) column. This became the point of merging the datasets. After merging and removing the null values, we were left with 101 regions of the world from 1990 to 2013 with 28,242 rows and 7 columns.

Null values of each column were dropped based on the skewness of the distribution plot. Seeing that the distribution plot of each column was either skewed to the left or right, rather than impute mean values or modal values, the missing values were dropped. Below is the distribution plot of each variable in the final merged dataset;

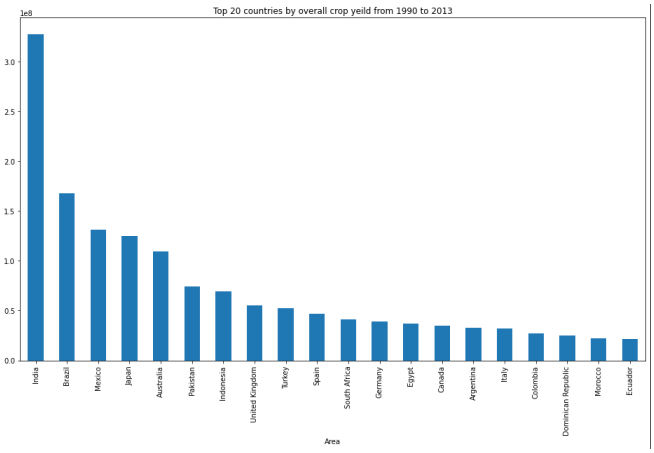




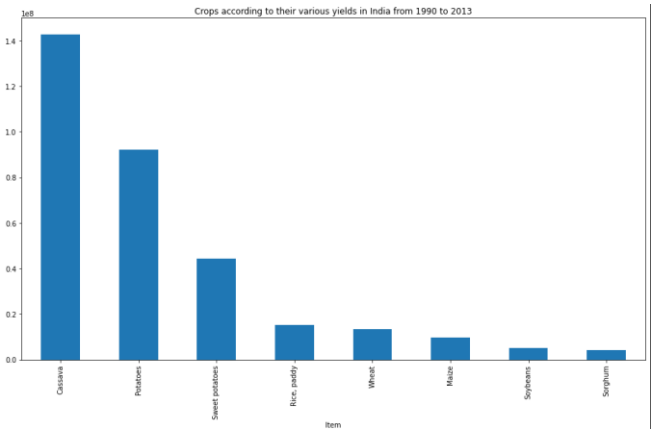


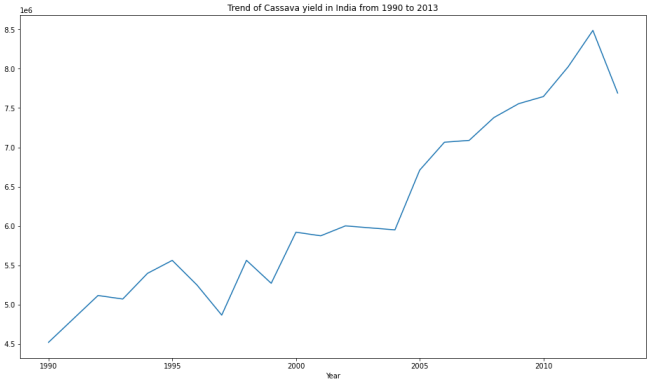


The data exploration provided insights as to regions with the highest yield of crops, crops with the highest yields in certain regions as shown in the visuals below;



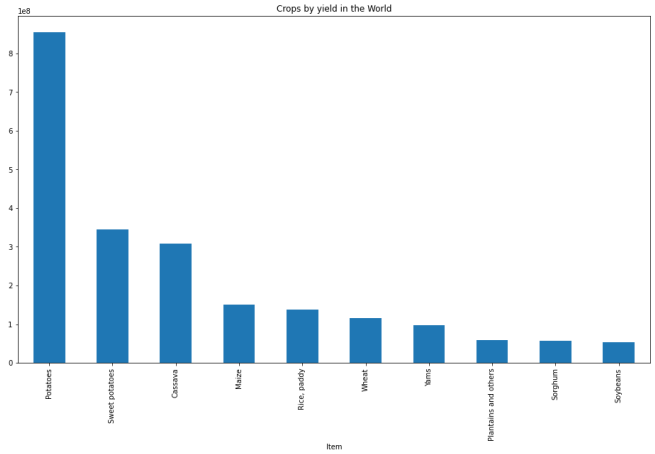
It can be observed that India produces more crops and achieve higher yields compared to other regions of the world. Driving further into India as a region, it was observed that Cassava had the highest yield and the trend of cassava yield from 1990 to 2013 was looked into as shown below.



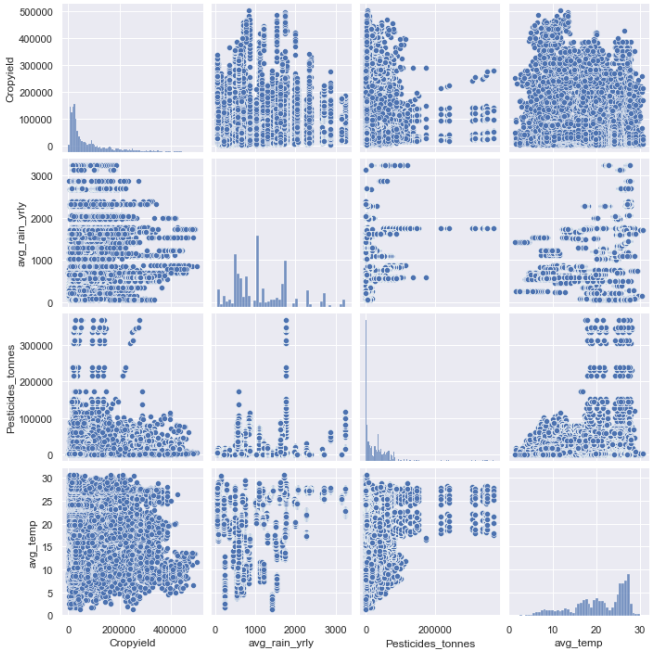


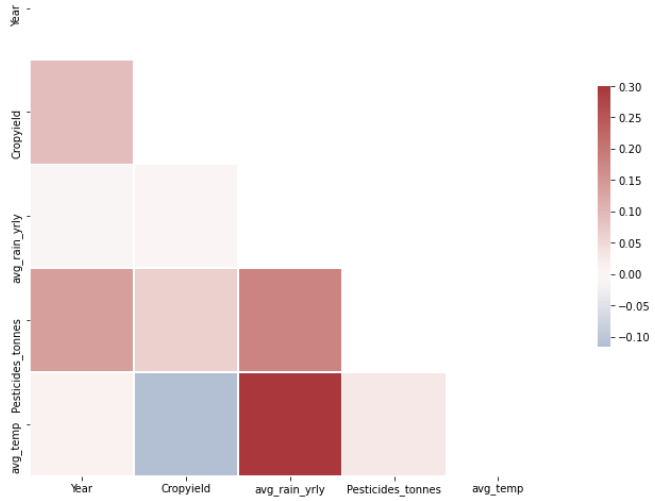
It can be observed that Cassava yields had been experiencing gradual increases over the years in India but seems to experience a decline towards 2013.

However, Potatoes achieved the highest yields across all 101 regions of the world as shown in the image below;



To ascertain the statistical measure of linear relationship between each of the variable. Correlation was implemented and visualized using heatmaps and also pair plots. This was done to ensure that none of the variables strongly influence one another.





Results show that none of the variable have any form of correlation.

V. Experiments

A. One Hot Encoding

Before going on to fit the models, the categorical variables within the dataset would have to be encoded to ensure that they are properly understood by the models to be utilized.
The categorical variables in the data set include Area and Item. They were encoded using one-hot encoding, leading to the creation of more columns within the dataset containing 1 and 0.

B. Define Feature and Target variables

After encoding, it was paramount to decide which column would be the target column where the prediction would take place and the features column, which determines the target column. The crop yield column was selected as the target column whilst all other columns were equated to the target variable. The features variable contained 28,242 inputs and 114 columns based on the encoded categorical variables.

C. Feature Scaling

Scaling the encoded variable could convert them to floats. Thus, the need for separating the numerical features before scaling as it was thought that it could affect the result. However, what was noted was that the same results were obtained even when the encoded variables were scaled. The essence of scaling was to ensure that all feature variables were brought to a similar level of magnitude, using the Minmax Scaler.

D. Train-Test Split

The data was split into training and testing in a 70:30 ratio using sklearn's train-test-split function. There were 19,769 entries in the training data and 8,473 entries in the test data.

VI. Machine Learning Models

A. Multiple Linear Regression

This happens to be the most frequently used ML model based on previous studies on the topic. However, it has been proven over time to produce low accuracies. It is included in this study for the sake of comparison.

B. Support Vector Regression

The Support Vector Regression model is a nonlinear development of Vapnik and Lerner's Generalized Portrait algorithm (1963). In its most basic form, the support vector technique aims to provide a linear equation f(x) = w,x> + b for a set of training data (x1,y1),...,(xm,ym) with w RN and b R.

C. K-Nearest Neighbour

This is another classifier that utilizes new inputs with vector y by examining the nearest neighbor in the training dataset points to y. In regression, the response value is derived as a weighted sum of all the nearest neighbours' replies, where the weight is inversely proportional to the distance from the input record (normalised Euclidean distance).

D. Random Forest

At training time, a random forest creates a huge number of decision trees and outputs the class that is the mean prediction (regression) of the individual trees.

E. XGBoost

XGBoost is an efficient, flexible, and portable distributed gradient boosting tool. It implements machine learning methods using the Gradient Boosting framework. XGBoost employs parallel tree boosting to swiftly and accurately solve data science problems..

F. Define Evaluation Metrics

The quality of each model was evaluated using the R square Score and Root Mean Square Error (RMSE) metrics. The R2 score is a measure of fit rather than forecast accuracy. The R square indicates how much variation in the model is explained by independent variables. If the score is close to one, the independent variables are responsible for all of the variance. The average difference between the model's projected and actual values in the dataset is the RMSE. As the RMSE drops, so does a model's capacity to "fit" a dataset.

VII. Results and Discussion

Results of the models employed are expressed in the table below:

Model	RMSE	R2 Score
Random Forest	13737.85	0.97
KNN	17659.44	0.96
SVR	20362.33	0.94
XGB Regressor	31295.94	0.86
Linear Regression	42728.03	0.75

According to the R2 score above, Random Forest had the best predictive accuracy (97%) and Linear regression had the weakest predictive accuracy (75%). According to the RMSE figures, Random Forest had the lowest RMSE and Linear Regression had the highest. Random Forest appears to be the best algorithm for this job.

When information is accessed, collected, processed, and used in machine learning initiatives, ethical problems are regularly highlighted. The data used for the study was obtained from a public source and does not comprise any sensitive information, signifying that we completed our data gathering stage without encountering any ethical problems.

Finally, there are no concerns regarding privacy or transparency with this project.

However, it's practically hard to claim such about our data preparation since some rows had null values which needed to be removed in order to favor the machine learning method. We had to remove approximately 2547 empty observations from the average temperature. But b Because removing these null values diminishes the data set's representation of that feature, it may raise ethical questions regarding bias. This can result in biased estimations and inaccurate conclusions, both of which can affect decision-making.

References

[1] A. T. M. S. Ahamed et al., "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015, pp. 1-6, doi: 10.1109/SNPD.2015.7176185.

[2] S. Khaki, L. Wang, and S. V. Archontoulis, 'A CNN-RNN Framework for Crop Yield Prediction', *Front. Plant Sci.*, vol. 10, p. 1750, Jan. 2020, doi: 10.3389/fpls.2019.01750.

[3] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, 'Predictive ability of machine learning methods for massive crop yield prediction', *Span J Agric Res*, vol. 12, no. 2, p. 313, Apr. 2014, doi: 10.5424/sjar/2014122-4439.

[4] T. van Klompenburg, A. Kassahun, and C. Catal, 'Crop yield prediction using machine learning: A systematic literature review', *Computers and Electronics in Agriculture*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[5] United Nations, 'World population projected to reach 9.7 billion by 2050', *World population projected to reach 9.7 billion by 2050*, Jul. 29, 2015. https://www.un.org/development/desa/en/news/population/2015-report.html#:~:text=The%20current%20world%20population%20of%207.3%20billion%20is,%E2%80%9CWorld%20Population%20Prospects%3A%20The%202015%20Revision%E2%80%9D%2C%20launched%20today. (accessed Apr. 11, 2022).

[6] L. Klerkx, E. Jakku, and P. Labarthe, 'A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda', *NJAS: Wageningen Journal of Life Sciences*, vol. 90–91, no. 1, pp. 1–16, Dec. 2019, doi: 10.1016/j.njas.2019.100315.

[7] J. H. Jeong *et al.*, 'Random Forests for Global and Regional Crop Yield Predictions', *PLoS ONE*, vol. 11, no. 6, p. e0156571, Jun. 2016, doi: 10.1371/journal.pone.0156571.

[8] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, 'Predictive ability of machine learning methods for massive crop yield prediction', *Span J Agric Res*, vol. 12, no. 2, p. 313, Apr. 2014, doi: 10.5424/sjar/2014122-4439.

[9] R. Beulah, 'A Survey on Different Data Mining Techniques for Crop Yield Prediction', *ijcse*, vol. 7, no. 1, pp. 738–744, Jan. 2019, doi: 10.26438/ijcse/v7i1.738744.

[10] Z. Chu and J. Yu, 'An end-to-end model for rice yield prediction using deep learning fusion', *Computers and Electronics in Agriculture*, vol. 174, p. 105471, Jul. 2020, doi: 10.1016/j.compag.2020.105471.

[11] M. Fernando, F. Cèsar, N. David, and H. José, 'Missing the missing values: The ugly duckling of fairness in machine learning', *Int J Intell Syst*, vol. 36, no. 7, pp. 3217–3258, Jul. 2021, doi: 10.1002/int.22415.