Honeybees and Neonic Pesticide

İrem Ekmekçi
Middle East Technical University
Ankara, Turkey
irem.ekmekci97@gmail.com

***Abstract*— This paper presents the prediction of honey yield per colony in USA with different regression models such as Artificial Neural Network, Support Vector Machine, Random Forest, Gradient Boosting, and XGBoost. R-Studio is used for this study. Preprocessed data is applied before constructing a model to predict honey yield per colony. The outlier analysis, data cleaning for explicit error and min-max normalization are conducted before making the prediction. The predictive performances of constructed models are evaluated by the root mean square error (RMSE) and mean absolute error (MAE).**

Keywords—Multiple Linear Regression, SVM, Artificial Neural Network, Random Forest, XGBoost

I. INTRODUCTION

In 2006, global concern was raised over the rapid decline in the honeybee population, an integral component to American honey agriculture. Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse. Speculation to the cause of this disorder points to hive diseases and pesticides harming the pollinators, though no overall consensus has been reached. Twelve years later, some industries are observing recovery but the American honey industry is still largely struggling. The U.S. used to locally produce over half the honey it consumes per year. Now, honey mostly comes from overseas, with 350 of the 400 million pounds of honey consumed every year originating from imports. This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012.

In this project, the main goal is predicting honey yield per colony using several regression model. The root mean squared error (RMSE) and mean absolute error (MAE) of each method are calculated and compared to identify their performances.

II. LITERATURE REVIEW

There are lots of researches on prediction of the amount of honey yield per colony. Tsukamoto's fuzzy inference system (FIS) method is used in [1] and, due to narrow data suggested data is not applicable for general. Another research about prediction of honey yield per colony in Turkey in [2] is with CART. Honey yield forecast is done in [3] by using Radial Basis Functions.

III. METHODOLOGY

A. Dataset

This data set is a combination of the data sets from The National Agricultural Statistics Service (NASS) and The United States Geological Survey (USGS). Moreover, it is taken from Kaggle. There are 825 observations with 16 variables in the data set. The data has 384 missing values. The variable description is given below.

- state: First two letter of the state - categorical
- numcol: Number of honey producing colonies - discrete
- yieldpercol: Honey yield per colony. Unit is pounds - discrete
- totalprod: Total production (numcol x yieldpercol). Unit is pounds - discrete
- stocks: Refers to stocks held by producers. Unit is pounds - discrete
- priceperlb: Refers to average price per pound based on expanded sales. Unit is dollars. - continuous
- prodvalue: Value of production (totalprod x priceperlb). Unit is dollars. - discrete
- year - discrete
- StateName: Full name of the state - categorical
- Region: Midwest/Northeast/South/West - categorical
- nCLOTHIANIDIN: The amount in kg of CLOTHIANIDIN applied - continuous
- nIMIDACLOPRID: The amount in kg of IMIDACLOPRID applied - continuous
- nTHIAMETHOXAM: The amount in kg of THIAMETHOXAM applied - continuous
- nACETAMIPRID: The amount in kg of ACETAMIPRID applied - continuous
- nAllNeonic: The amount in kg of all Neonics applied = (nCLOTHIANIDIN + nIMIDACLOPRID + nTHIAMETHOXAM + nACETAMIPRID + nTHIACLOPRID) - continuous

B. Descriptive Statistics

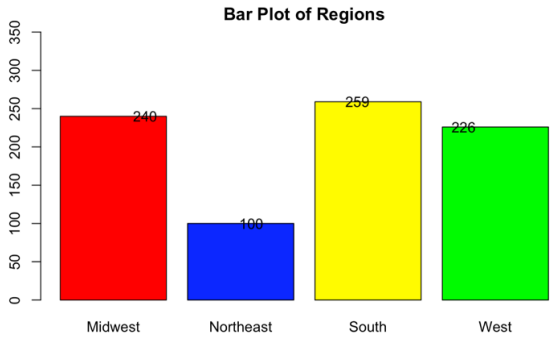
Descriptive statistics table are shown below. It gives the summary statistics of some numerical variables.

	num-col	yi-eld-per-col	pri-ce-perlb	prodva-lue	nACE-TA-MIP-RID
Min.	2000	19	0.490	162000	0.0
1st Qu.	9000	46	1.100	924000	0.0
Medi-an	26000	57	1.520	2132000	16.0
Mean	61890	60.22	1.758	5608421	729.0
3rd Qu.	65000	72	2.090	5599000	349.1
Max.	510000	136	7.860	83859000	36480.3
NA's	0	0	0	0	64

Table 1 Descriptive Statistical Summary of Numerical Data

As an example, the average amount of ACE-TAMIPRID applied is 729.0. The minimum amount of ACETAMIPRID applied is 0.0, while the maximum is 36480.3. The half of amount of ACETAMIPRID applied is below or above 16.0. 25% of amount of ACETAMIP-RID applied is below 0.0 and above 349.1. In addition, this variable has 64 missing values.

Figure 1 Bar Plot of Regions



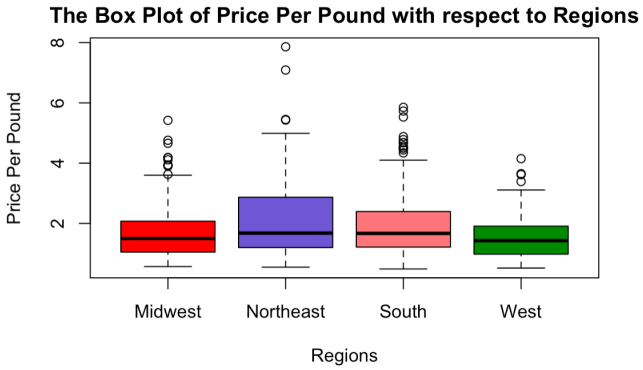
According to bar plot, South is the most common region.

C. *Explanatory Data Analysis*

In this part, four research questions has been created and solved by using suitable statistical methods.

RQ1.Is the average price per pound the same for all regions?

To see whether the regions have the same average price per pound or not, One-Way ANOVA or Kruskal-Wallis should be used. However, let's check it with a box plot.



It is seen that the average values of the price per pound of each region represented by the median are different for each case.

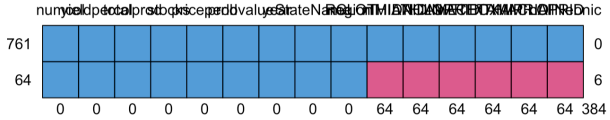
Since the populations are independent from each other, without assuming normality and equality of variances, the populations can be compared with Kruskal-Wallis Test. Kruskal-Wallis test is nonparametric version of the One-Way ANOVA.

```
Kruskal-Wallis Test (alpha = 0.05)
------------------------------------------------------------
data : priceperlb and Region

statistic  : 22.14119
parameter  : 3
p.value    : 6.096403e-05

Result     : Difference is statistically significant.
------------------------------------------------------------
```

Kruskal- Wallis test shows that the average price per pound for each region are statistically different .D.

D. *Missingness*

Data is often described in accordance to the reasons for the missing data. In this data set, there are missing values created by the mechanism of missing completely at random.



The plot shows that there are total 384 missing observations. Also, it shows the number of missing observation that each variable has. However, this missing plot is not clear.

By using “mice” package NA’s are imputed.

Before imputation:

```
    state         numcol        yieldpercol       totalprod
AL      : 20   Min.   :  2000   Min.   : 19.00   Min.   :   84000
AR      : 20   1st Qu.:  9000   1st Qu.: 46.00   1st Qu.:  470000
AZ      : 20   Median : 26000   Median : 57.00   Median : 1500000
CA      : 20   Mean   : 61890   Mean   : 60.22   Mean   : 4117353
CO      : 20   3rd Qu.: 65000   3rd Qu.: 72.00   3rd Qu.: 4032000
FL      : 20   Max.   :510000   Max.   :136.00   Max.   :46410000
(Other):705
    stocks           priceperlb      prodvalue            year
Min.   :    8000   Min.   :0.490   Min.   :  162000   Min.   :1998
1st Qu.:  117000   1st Qu.:1.100   1st Qu.:  924000   1st Qu.:2002
Median :  383000   Median :1.520   Median : 2132000   Median :2007
Mean   : 1233493   Mean   :1.758   Mean   : 5608421   Mean   :2007
3rd Qu.: 1361000   3rd Qu.:2.090   3rd Qu.: 5599000   3rd Qu.:2012
Max.   :13800000   Max.   :7.860   Max.   :83859000   Max.   :2017

     StateName         Region    nCLOTHIANIDIN      nIMIDACLOPRID
Alabama   : 20   Midwest  :240   Min.   :     0.0   Min.   :     3.2
Arizona   : 20   Northeast:100   1st Qu.:     0.0   1st Qu.:   937.5
Arkansas  : 20   South    :259   Median :   336.7   Median :  3698.9
California: 20   West     :226   Mean   : 10890.9   Mean   : 10019.4
Colorado  : 20                   3rd Qu.:  6569.3   3rd Qu.: 10588.6
Florida   : 20                   Max.   :278498.8   Max.   :150569.3
(Other)   :705                   NA's   :64         NA's   :64
nTHIAMETHOXAM      nACETAMIPRID      nTHIACLOPRID      nAllNeonic
Min.   :    0.0   Min.   :    0.0   Min.   :   0.0   Min.   :     3.2
1st Qu.:   29.7   1st Qu.:    0.0   1st Qu.:   0.0   1st Qu.:  1631.8
Median : 1156.3   Median :   16.0   Median :   0.0   Median :  8402.1
Mean   : 6225.1   Mean   :  729.0   Mean   : 120.5   Mean   : 27984.9
3rd Qu.: 7777.0   3rd Qu.:  349.1   3rd Qu.:   0.0   3rd Qu.: 33508.2
Max.   :64834.6   Max.   :36480.3   Max.   :4273.2   Max.   :403011.6
NA's   :64        NA's   :64        NA's   :64       NA's   :64
```

After imputation:

```
    state         numcol        yieldpercol       totalprod
AL      : 20   Min.   :  2000   Min.   : 19.00   Min.   :   84000
AR      : 20   1st Qu.:  9000   1st Qu.: 46.00   1st Qu.:  470000
AZ      : 20   Median : 26000   Median : 57.00   Median : 1500000
CA      : 20   Mean   : 61890   Mean   : 60.22   Mean   : 4117353
CO      : 20   3rd Qu.: 65000   3rd Qu.: 72.00   3rd Qu.: 4032000
FL      : 20   Max.   :510000   Max.   :136.00   Max.   :46410000
(Other):705
    stocks           priceperlb      prodvalue            year
Min.   :    8000   Min.   :0.490   Min.   :  162000   Min.   :1998
1st Qu.:  117000   1st Qu.:1.100   1st Qu.:  924000   1st Qu.:2002
Median :  383000   Median :1.520   Median : 2132000   Median :2007
Mean   : 1233493   Mean   :1.758   Mean   : 5608421   Mean   :2007
3rd Qu.: 1361000   3rd Qu.:2.090   3rd Qu.: 5599000   3rd Qu.:2012
Max.   :13800000   Max.   :7.860   Max.   :83859000   Max.   :2017

     StateName         Region    nCLOTHIANIDIN      nIMIDACLOPRID
Alabama   : 20   Midwest  :240   Min.   :     0.0   Min.   :     3.2
Arizona   : 20   Northeast:100   1st Qu.:     0.0   1st Qu.:  1101.0
Arkansas  : 20   South    :259   Median :   443.8   Median :  3964.6
California: 20   West     :226   Mean   : 11847.0   Mean   : 10712.6
Colorado  : 20                   3rd Qu.:  8067.6   3rd Qu.: 11217.8
Florida   : 20                   Max.   :278498.8   Max.   :150569.3
(Other)   :705
nTHIAMETHOXAM    nACETAMIPRID      nTHIACLOPRID      nAllNeonic
Min.   :    0   Min.   :    0.0   Min.   :   0.0   Min.   :     3.2
1st Qu.:   45   1st Qu.:    0.0   1st Qu.:   0.0   1st Qu.:  1863.2
Median : 1423   Median :   20.8   Median :   0.0   Median : 10063.1
Mean   : 6614   Mean   :  758.0   Mean   : 119.3   Mean   : 29915.1
3rd Qu.: 8552   3rd Qu.:  373.0   3rd Qu.:   0.0   3rd Qu.: 35930.5
Max.   :64835   Max.   :36480.3   Max.   :4273.2   Max.   :403011.6
```

E. Modelling

After the data imputation, the data set is ready to conduct a model to predict honey yield per colony.

Firstly, the normality of the variable yieldpercol is checked, and the variable is normalized with the Tukey Ladder Transformation. Then, before constructing a model, the data is divided into two parts which are train data and test data. Dividing the data set into train and test is called Cross-Validation. Cross-Validation is used to test the suitability of the model when there are new observations about the variables.

Therefore, 662 observations are used as a train data to construct a model. Also, 163 observations are accepted as a test data..

1. Multiple Linear Regression

Multiple linear is a statistical model that shows the linear relationship between response and independent variable/s.

The insignificant variables are eliminated using forward elimination method. Hence, the final model is fitted. The output of the final model is given below.

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	1,54E+03	4,776E+00	322.450	< 2e-16
numcol	-1,569E-04	4,817E-05	-3.256	0.00119
priceperlb	-1,939E+01	1,756E+00	-11.040	< 2e-16
prodvalue	2,422E-06	4,335E-07	5.587	3.40e-08
Region2	-2,923E+01	5,699E+00	-5.128	3.86e-07
Region3	-1,867E+00	4,322E+00	-432	0.66588
Region4	-1,854E+01	4,535E+00	-4.088	4.89e-05
nACETAMIPRID	-1,658E-03	6,978E-04	-2.376	0.01779

Residual standard error: 0.04171 on 654 degrees of freedom
Multiple R-squared: 0.274, Adjusted R-squared: 0.2663
F-statistic: 35.27 on 7 and 654 DF, p-value: < 2.2e-16
Table The Results of Multiple Linear Regression

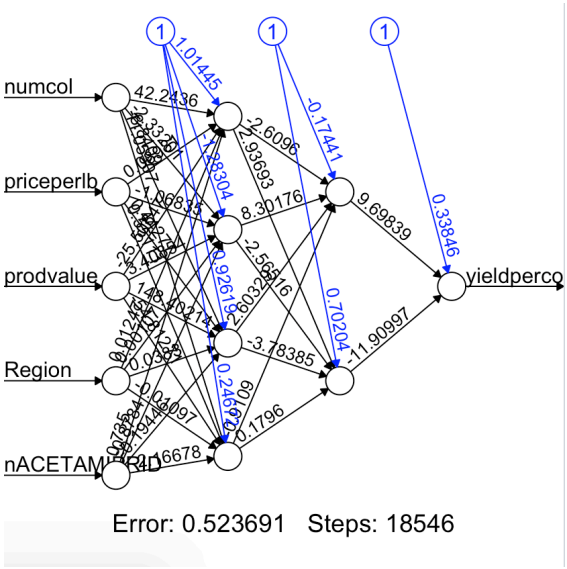
The model is significant since p-value is less than 0.05. It is clearly seen that only numcol has a positive effect on honey yield per colony. Other variables have negative effect on honey yield per colony.
About 26% of the variation in honey yield per colony can be explained by number of honey producing colonies, price per pound, value of production, region, the amount of ACETAMIPRID applied.

2. Artificial Neural Network

Artificial Neural Network is supervised learning algorithm that can be used for classification and regression problem.

In this method, it is tried to predict honey yield per colony by considering number of honey producing colonies, price per pound, value of production, region, and the amount of ACETAMIPRID applied.

The features are normalized after the determining them. The ANN model is constructed with two hidden layers. The first layer has 4 neurons, and the second layer has two neurons. The plot of the model is given below.



Figures Representation of Neural Network for Two Hidden Layers

3. Support Vector Machine

Support Vector Machine is also supervised learning algorithm that can be used for classification and regression problem.

In this method, it is tried to predict honey yield per colony by considering number of honey producing colonies, price per pound, value of production, region, and the amount of ACETAMIPRID applied.

After deciding features, SVM parameters are tuned. As a result, cost and gamma parameters are found 100 and 0.1, respectively.

4. Random Forest

Random forest is an another supervised learning based on tree algorithms which is applicable for both regression and classification problems.

In this method, it is tried to predict honey yield per colony by considering number of honey producing colonies, price per pound, value of production, region, and the amount of ACETAMIPRID applied.

The number of trees providing the lowest error rate is found which is 490 trees providing an average honey yield per colony error of 0.03255869.

5. Gradient Boosting Algorithm

Gradient Boosting Algorithm is also supervised learning algorithm that can be used for classification and regression problem.

In this method, it is tried to predict honey yield per colony by considering number of honey producing colonies, price per pound, value of production, region, and the amount of ACETAMIPRID applied.

The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.

6. XGBoost

XGBoost is a machine learning method based on tree algorithm and can be used for both regression and classification problems.

In this method, it is tried to predict honey yield per colony by considering number of honey producing colonies, price per pound, value of production, region, and the amount of ACETAMIPRID applied.

7.Error Control Methods

In this work, RMSE and MAE are used to calculate errors.

- Root Mean-Squared Error (RMSE)
- Mean Absolute Error (MAE)

IV. RESULTS

In this part, the results are expressed for following classification models;

i. Multiple Linear Regression
ii. Artificial Neural Network
iii. Support Vector Machine
iv. Random Forest
v. Gradient Boosting Algorithm
vi. XGBoost

In Table , RMSE and MAE are compared. According to table the table, best prediction model is XGBoost

	RMSE	**MAE**
Multiple Linear Regression	0.03806025	0.02998847
Artificial Neural Network	0.03977624	0.02876655
Support Vector Machine	0.04667056	0.03839639
Random Forest	0.02958764	0.02336891
Gradient Boosting	0.02251069	0.01745183
XGBoost	0.01637439	0.01259167

V. CONCLUSION

In this project, exploratory data analysis like the graphical techniques and descriptive statistics are conducted. Then, the data cleaning for missing values is applied to the data in order to make the quality of the data better. Finally, honey yield per colony is tried to predict by using several methods. Their results are shown in the previous chapter. According to provided data, it is observed that number of honey producing colonies, price per pound, value of production, region, the amount of ACETAMIPRID applied are the most efficient factors on honey yield per colony. As a result, it is found that XGBoost is the best method to predict honey yield per colony.

VI. REFERENCES

[1] T. Hastono, A. J. Santoso and Pranowo, "Honey yield prediction using Tsukamoto fuzzy inference system," *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Yogyakarta, 2017, pp. 1-6, doi: 10.1109/EECSI.2017.8239150.

[2] Karadas, Koksal & Kadirhanoğulları, Ibrahim. (2017). Predicting Honey Production using Data Mining and Artificial Neural Network Algorithms in Apiculture. Pakistan Journal of Zoology. 49. 1611-1619. 10.17582/journal.pjz/2017.49.5.1611.1619.

[3] H. Rocha and J. Dias, "Honey Yield Forecast Using Radial Basis Functions", CeBER and Faculdade de Economia, Universidade de Coimbra, 3004-512 Coimbra, Portugal.