The bees' needs: A supervised model for predicting honeybee hive mortality risk based on disease status, productivity, geography, temperature, management, and species.

Timothy Lam*, Samantha Share*

*Department of Integrative Biology

University of Guelph, Guelph, Ontario, N1G 2W1, Canada

April 21, 2023

Abstract

Honeybee hive mortality is important for beekeepers in terms of their livelihood and crop production and should be monitored to reduce hive loss. This project comprised of running four algorithms, random forest, decision trees, support vector machines, and gradient boosting machines, to predict seasonal mortality based on previously known variables including the environment surrounding the apiary, instances of disease, and how many queens and swarms were bought. Feature selection and recursive feature elimination showed that the most accurate model was using random forest and including all the variables. A common issue found in all the learning curves was underfitting prior to hyperparameter optimization, however, this was partially corrected after the tuning. Mean squared errors were found to be quite high among certain models, and future considerations would allow for more complex methods at reducing these values. A user input program called the colony mortality prediction program was implemented to allow beekeepers to input known data on their hives to be given an output of what their seasonal mortality will be.

Key words:

Machine learning, honey bees, colony collapse, colony mortality

List of Abbreviations

CCD = Colony collapse disorder

CART = Classification and regression trees

SVM = Support vector machines

GBM = gradient boosting machines

DT = Decision trees

RF = Random Forest

EPILOBEE = epidemiological surveillance program on honeybee colony mortality

MSE = Mean squared error:

1. Introduction

Honeybee Background

As one of the most important pollinators responsible for a large sum of the world's food crops, it is vital that honeybee hive mortality is monitored. Beekeepers rely on honeybees for economic stability, crop production, and producing by-products such as beeswax and royal jelly to make cosmetic products and candles (Sillman et al., 2021). In 2019, 80 million pounds of honey were produced in Canada, which was integral to the almost 5 billion dollar contribution to the Canadian economy (Bixby et al., 2021). Honeybees play a crucial role in maintaining the health of many ecosystems by pollinating plant species to provide habitat and food for other wildlife (Hung et al., 2018). 87.5% of flowering plant species are believed to be pollinated by animals, and honeybees are successful pollinators in either their native or introduced environment (Hung et al., 2018). The pollination of medicinal plants and herbs are relied upon for treating diseases and illnesses (Durazzo et al., 2021). Many bee products act as a source of antioxidants that are believed to work against the effects of oxidative stress, making them vital to human use (Durazzo et al., 2021). The decline of the global honeybee population has become a larger problem in recent years and is due to pesticide use, climate change, and the spread of

diseases ((Cornelissen et al., 2019; Chambó, 2016))).

Colony Collapse Disorder (CCD) is described as the sudden disappearance of adult bees from their colonies that contain sufficient brood and food levels with no indicator of parasites (Williams et al., 2010). There is no known cause for CCD, and it is a leading cause of honeybee hive mortality. CCD is not well understood by scientists, and this lack of answers causes beekeeper's rising concern over their hives. There are many factors involved in this population decline that can give beekeepers a better idea of what to expect. Pesticide exposure and viral infections are both associated with weakened colonies (Münch & Amdam, 2013). Queen failure, starvation, climate change, and varroa mites are also commonly linked to honeybee hive loss and better understanding and documentation of these events could allow for beekeepers to predict their hive loss (James & Li, 2012).

Algorithm introduction

Decision trees are another algorithm that are built by looking at training examples with known class labels (Kingsford & Salzburg. 2008). Decision trees are used for classifying tumours or making technological and business decisions due to their ability to break down complicated data (Kingsford & Salzburg. 2008). Data is classified by various questions about the features, and each question is in a node with internal nodes associated with the possible answers (Kingsford & Salzburg. 2008). Decision trees have the advantage of being easier to interpret as the data is categorized into simple questions, and they are flexible with the inputs (Kingsford & Salzburg. 2008).

Random forest is a supervised machine learning algorithm that is a combination of Classification and Regression Trees (CART) that is trained on a dataset (Sarica et al., 2017). When a tree is built, a collection of bootstraps is created without records included in the original

data as out-of-bag samples to act as a test set (Sarica et al., 2017). When classifying new data, every CART tree votes for a class to allow the forest to predict the class with the most votes among all the trees (Sarica et al., 2017). A random forest algorithm is used in many settings as it is useful for both classification and regression data and has high accuracy and scalability without causing high amounts of overfitting (Sarica et al., 2017).

Support vector machines (SVMs) are useful for both classification and regression algorithms and are a good tool for pattern recognition (Cervantes et al., 2019). This algorithm transforms the data to find the boundary between the potential outputs to separate it (Cervantes et al., 2019). SVMs are not as successful for unbalanced data and can take more time due to the complexity, however they are very useful for large data sets in high dimensional space (Cervantes et al., 2019).

Gradient Boosting Machines (GBMs) combine several weak models to improve the predictions. GBMs iteratively add new models to the ensemble to fix the mistakes made by the weaker models, and they train the new models on the residual errors (Natekin & Knoll, 2013). This algorithm finds the optimal parameters using the gradient descent and a cost function measures the difference between the actual and predicted values of the target output (Natekin & Knoll, 2013). Boosting is also used to weigh the contribution of each model in the ensemble (Natekin & Knoll, 2013). GBMs are useful for many ranges of data types and are a reliable tool for machine learning.

2. Materials and Methods

EPILOBEE dataset:

To conduct our analysis, we obtained the epidemiological surveillance program on

honeybee colony mortality (EPILOBEE) dataset from: https://zenodo.org/record/269636#.ZD9akXbMI2w. This dataset was constructed between 2012 and consists of a variety of metrics regarding the health of the honeybee colonies. The features in this dataset can be broken into three general categories: Information on beekeepers, information on bee and hive health, and information on geography. For a complete summary of the features in this dataset see table 1. Prior to processing, this dataset had 4759 observations.

Data processing:

The first step of data processing was the addition of a 'Mean_Temperature' column. This was accomplished by extracting the mean temperatures for each of the countries in the EPILOBEE dataset for the period between September 2012 to September 2014 from a NASA dataset of mean temperatures in europe (https://data.open-power-system-data.org/weather_data/2020-09-16).

Next, categorical variables in the EPILOBEE dataset were converted to integers and all rows with NAs or missing values were removed. Finally, the ID and program columns were removed as it was solely important in data collection and would not have predictive value in colony mortality. After removing rows with NAs and missing values, we were left with 4732 observations.

Problem definition:

We utilized all features in our processed EPILOBEE dataset to predict the percent colony mortality of an apiary during the beekeeping season. The response variable name in the dataset is "seasonal_mortality". The seasonal mortality is a continuous variable, thus this will be a regression problem.

Recursive feature elimination and model training:

Recursive feature elimination was performed using the four algorithms mentioned in the introduction; SVM, DT, RF, and GBM to identify subsets of sizes 5, 10, 20 and 30. Next, the data was split into training and testing sets with a 75/25 split. Each of the aforementioned models were trained on their respective subsets and the entire dataset of 37 features.

Hyperparameter optimization

Next, we conducted hyperparameter optimization on each of the models. RandomizedSearchCV and GridSearchCV from Scikit-Learn were utilized for hyperparameter optimization. See table 2 for a summary of parameters used for each algorithm. Parameters for SVM were limited due to the computational complexity of the algorithm.

Testing and saving models:

Model predictions were tested against the testing set, with mean squared error (MSE) as the metric of comparison between models. A total of 20 models were created and saved as .sav files. In addition to saving models, a list of the features utilized in each of the models was saved as .txt files.

Feature Selection:

The data from the processed and cleaned file was split into X and Y arrays for the input and output variables. SelectKBest was used to select the top 10 features based on their ANOVA F-value scores. Figure 4 is a bar plot of the top ten variables with the highest scores.

Colony mortality prediction program:

Finally, we developed a python program to query users for a variety of features corresponding to the features in the EPILOBEE dataset. Next, their responses are compiled in a Pandas dataframe and the best performing model that was trained on the features that were inputted by the user is utilized to make a prediction. This maximizes the likelihood that a user

who has fewer metrics is still able to obtain a precise prediction of honeybee hive mortality. An internal metric is used to set the highest acceptable MSE for the models used in prediction. As a default, our highest acceptable score of MSE is 144, corresponding to an error rate of ± 12%. If the user inputs insufficient metrics, or the metrics match to a model with an MSE score greater than 144, it will output a message informing the user that they must input more information and suggest specific metrics that the user should obtain information on to obtain an accurate prediction.

3. Results

Feature importance and dataset overview:

A positive correlation was found between two of the features, the number of queens bought and the number of swarms bought, in Figure 1. The rest of the correlations were close to zero or negative. Figure 2 shows a class imbalance in the existing seasonal mortality data with over 4000 entries of zero. The initial plots showed potential variables associated with disease are related to seasonal mortality, as shown in Figure 3. Figure 4 shows that the top ten features impacting seasonal mortality are Activity, Bee_population_size, Country, Apiary_Size, Chronic_Depop, ClinSign_Brood, H_Rate_HoneyMortality, EuropeanFoulbroodV2, Merger, and Mean_Temperature.

Learning curves:

Figures 5,6,7,and 8 show the best learning curve figures for each algorithm, which was five features for each. Figure 5 shows that the random forest learning curve with five features had underfitting that was corrected by hyperparameter optimization. Figures 6 and 7 do not show as much of a change from optimization, but figure 8 does. The learning curves were made for

every model and showed similar results of correcting underfitting, however the plots for five features showed the most change.

Performance of optimized models:

Table 3 shows the testing MSE of each of the models. The lowest overall MSE was RF with a subset of 30 features, followed closely by the same algorithm with the full set of features in the EPILOBEE dataset. With the smallest subset of 5 features the DT algorithm performed with the lowest MSE. With the larger sets of 20, 30, and 37 features, the DT algorithm and GBM algorithm performed with very similar MSEs. The SVM algorithm performed the worst compared to all other models with a MSE of greater than 120 regardless of the size of the feature set.

4. Discussion

Analysis of model performance:

SVMs performed with the highest MSE compared to all other algorithms and their performance was only minorly improved via hyperparameter optimization (Table 3, Figure 7). This may be in part due to the exceeding computational complexity of SVMs forced the authors to reduce the number of parameters used in hyperparameter optimization leading to reduced precision of the SVM-based models overall. Figure 9 shows that hyperparameter optimization had little impact on the training scores of the SVM-based models. It is also likely that SVMs are not appropriate for this dataset. Both tree-based methods performed very well. This was unsurprising, as DT and RF rely on partitioning the data with binary splits which is made easy by the exclusively categorical predictor variables. The performance of the tree-based models was further improved by hyperparameter optimization (Figures 10,11). Figure 12 shows that

hyperparameter optimization of the GBM-based models resulted in insignificant improvements. This may indicate in the future we could increase the hyperparameter space to improve the efficacy of hyperparameter optimization.

The MSE of our best performing models corresponds to an error rate of ± ~10% for the prediction of colony mortality by our models. We deemed this error rate sufficiently low for the application of these models into our colony mortality prediction program but believe it could be reduced further by making improvements to the EPILOBEE dataset and addressing the underfitting that was common in several of the models.

Learning Curves:

The learning curves showed that hyperparameter optimization was highly successful at accounting for some of the underfitting shown in many of the pre-optimization learning curves. The learning curves with five features, Figures 5 through 8, had the most improvement after optimization. This is likely because adding more features makes the model more complex. Figure 7 showed that the SVM outputs had a horizontal cross-validation score, indicating this is likely not a good fit. Even though the learning curves for fewer features showed more promising results, including more features in the algorithm was shown to be more successful at predicting seasonal mortality of honeybee hives.

Modifications to EPILOBEE dataset:

To further increase the efficacy of the colony mortality predicting program, we believe several changes to the EPILOBEE dataset could be made. If it is possible, a crucial first step would be replacing several of the categorical features with continuous features. For example, several measurement variables, like apiary size or number of swarms produced, are naturally continuous variables but have been binned into categories by the creators of the dataset.

Furthermore, the production column has been compressed in a way that does not reflect the complexity of the production of the individual apiaries, specifically, for apiaries that produce products other than only honey, it is unknown what other products they produce.

In addition to changes to the existing variables, we believe the addition of several variables may increase the predictive ability of the models. Firstly, there are several pathogens of honeybees that were not included in the dataset including: Acute bee paralysis virus, Israeli acute paralysis virus, Kashmir bee virus, Black queen cell virus, Sacbrood virus and chalkbrood (Chen and Siede, 2007; Ye et al., 2021). These pathogens have been associated with colony collapse disorder and likely have an impact on colony mortality (Chen and Siede, 2007; Ye et al., 2021). Next, focus could be placed on adding more information on the specific pesticide exposure for each of the apiaries. Currently, there is no feature for presence of any pesticide in the EPILOBEE dataset and pesticide exposure can only be inferred from the countries and environment the apiaries are in. Specific testing for pesticides at the sources of exposure to honeybees (nectar, pollen, and water sources) could be implemented to quantify the presence of specific pesticides at each apiary (Chambó, 2016). A further stressor on honeybee health is changing climate (Reddy et al., 2012; Cornelissen et al., 2019; Abou-Shaara et al., 2017). Features that could be added to reflect a changing climate include the difference between the current average temperature compared to historical average temperatures and air quality reports from the regions where the apiaries are located.

Addressing model underfitting:

As observed in figure 13, underfitting was a common occurrence with our models that were trained on the larger number of features. We believe that the aforementioned modifications to the EPILOBEE dataset may aid in reducing the underfitting we noticed with several models.

However, a larger, more complex dataset may require more complex algorithms to prevent similar underfitting. In the future, we hope to evaluate the potential of extreme gradient boosting and artificial neural networks to determine if these algorithms are appropriate for our dataset.

Colony mortality prediction program:

Currently, the colony mortality prediction program is in a very preliminary state and there are several updates that could increase its usability. As of April 21, 2023 the user-input section queries the users to input a number corresponding to the bin for that specific category. In the future, we hope to improve this so the user is queried only for a value and encoding occurs entirely on the back-end. Furthermore, in a future version we plan to implement functionality that suggests treatments or factors that should be closely monitored in cases where users whose inputs yield predictions with high risk of colony mortality. Finally, it is our aim to integrate this program into an app to ensure accessibility for beekeepers at all levels.

5. Conclusion:

To address the phenomenon of colony collapse disorder, we attempted to develop a ML approach to predict the risk of colony mortality in a given apiary. Using a variety of subsets of the modified EPILOBEE dataset, we were able to train and test several models with the RF, DT, SVM, and GBM algorithms. These models performed with varying levels of precision in predicting colony mortality. These models were used to construct a program for predicting colony mortality of any given apiary. While we were pleased with the performance of several of the models, we believe several improvements could be made to further increase their performance. This could include improving the EPILOBEE dataset by modifying existing variables and adding new variables that may have value in predicting colony mortality. Additionally, experimentation with more complex algorithms and potentially exploring deep

learning approaches may improve MSE and reduce underfitting. We hope that with improvements, our colony mortality predicting program can be implemented in apiaries around the world to predict the risk of colony mortality.

Acknowledgements:

The authors would like to thank Antoine Jacques and their colleagues at the European Food Safety Agency for their work in building the EPILOBEE dataset. The authors would also like to thank Dr. Dan Tulpan (University of Guelph, department of Animal Biosciences) for his guidance and feedback while completing this project.

Authors Contributions:

All authors improved and contributed to the editing of the manuscript. All authors read and approved the final manuscript.

Disclosures:

The authors declare no real or perceived conflicts of interest.

Literature cited:

Abou-Shaara, H. F., A. A. Owayss, Y. Y. Ibrahim, and N. K. Basuny. 2017. A review of impacts of temperature and relative humidity on various activities of honey bees. Insect. Soc. 64:455–463. doi:10.1007/s00040-017-0573-8.

Antoine, J., L. Marion, R.-C. Magali, S. Mathilde, B. Stéphanie, H. Pascal, and C. Marie-Pierre. 2016. Honey bee Seasonal mortality 2012-2014 - Epilobee analysis. doi:10.5281/zenodo.269636. Available from: https://zenodo.org/record/269636

Bixby, M. E. F., M. Polinsky, R. Scarlett, H. Higo, J. Common, S. E. Hoover, L. J. Foster, A. Zayed, M. Cunningham, and M. M. Guarna. 2021. Impacts of COVID-19 on Canadian Beekeeping: Survey Results and a Profitability Analysis. D. Tarpy, editor. Journal of Economic Entomology. 114:2245–2254. doi:10.1093/jee/toab180.

Cervantes, J., F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 408:189–215. doi:10.1016/j.neucom.2019.10.118.

Chambó, E. 2016. Beekeeping and Bee Conservation: Advances in Research. BoD – Books on Demand.

Chen, Y. P., and R. Siede. 2007. Honey Bee Viruses. In: Advances in Virus Research. Vol. 70. Academic Press. p. 33–80. Available from: https://www.sciencedirect.com/science/article/pii/S0065352707700027

Cornelissen, B., P. Neumann, and O. Schweiger. 2019. Global warming promotes biological invasion of a honey bee pest. Glob Chang Biol. 25:3642–3655. doi:10.1111/gcb.14791.

Durazzo, A., M. Lucarini, M. Plutino, G. Pignatti, I. K. Karabagias, E. Martinelli, E. B. Souto, A. Santini, and L. Lucini. 2021. Antioxidant Properties of Bee Products Derived from Medicinal Plants as Beekeeping Sources. Agriculture. 11:1136. doi:10.3390/agriculture11111136.

Hung, K.-L. J., J. M. Kingston, M. Albrecht, D. A. Holway, and J. R. Kohn. 2018. The worldwide importance of honey bees as pollinators in natural habitats. Proc. R. Soc. B. 285:20172140. doi:10.1098/rspb.2017.2140.

James, R. R., and Z. Li. 2012. From Silkworms to Bees. In: Insect Pathology. Elsevier. p. 425–459. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123849847000129

Kingsford, C., and S. L. Salzberg. 2008. What are decision trees? Nat Biotechnol. 26:1011–1013. doi:10.1038/nbt0908-1011.

Münch, D., and G. V. Amdam. 2013. Brain Aging and Performance Plasticity in Honeybees. In: Handbook of Behavioral Neuroscience. Vol. 22. Elsevier. p. 487–500. Available from: https://linkinghub.elsevier.com/retrieve/pii/B978012415823800037X

Pfenninger, S., and I. Staffell. 2020. Weather Data. doi:10.25832/WEATHER_DATA/2020-09-16. Available from: https://data.open-power-system-data.org/weather_data/2020-09-16

Reddy, P. V. R., A. Verghese, and V. V. Rajan. 2012. Potential impact of climate change on honeybees (Apis spp.) and their pollination services. 18.

Sarica, A., A. Cerasa, and A. Quattrone. 2017. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front. Aging Neurosci. 9:329. doi:10.3389/fnagi.2017.00329.

Sillman, J., V. Uusitalo, T. Tapanen, A. Salonen, R. Soukka, and H. Kahiluoto. 2021. Contribution of honeybees towards the net environmental benefits of food. Science of The Total Environment. 756:143880. doi:10.1016/j.scitotenv.2020.143880.

Williams, G. R., D. R. Tarpy, D. vanEngelsdorp, M.-P. Chauzat, D. L. Cox-Foster, K. S. Delaplane, P. Neumann, J. S. Pettis, R. E. L. Rogers, and D. Shutler. 2010. Colony Collapse Disorder in context. Bioessays. 32:845–846. doi:10.1002/bies.201000075.

Ye, M.-H., S.-H. Fan, X.-Y. Li, I. M. Tarequl, C.-X. Yan, W.-H. Wei, S.-M. Yang, and B. Zhou. 2021. Microbiota dysbiosis in honeybee (Apis mellifera L.) larvae infected with brood diseases and foraging bees exposed to agrochemicals. R Soc Open Sci. 8:201805. doi:10.1098/rsos.201805.

Figures

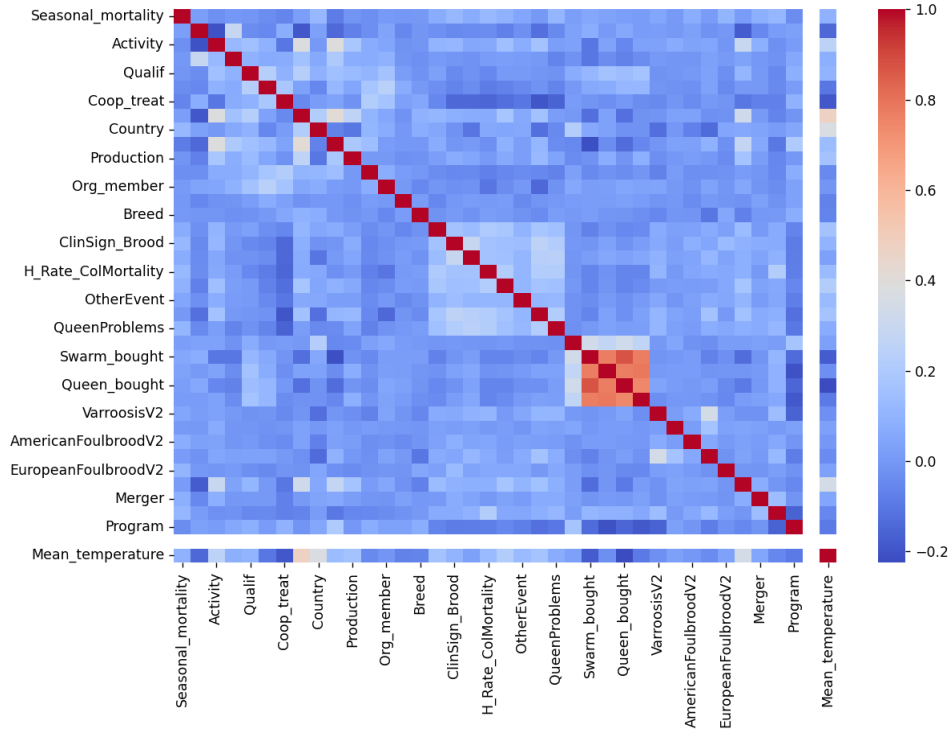

Figure 1. Heatmap of correlation matrix of the features in the bee dataset.

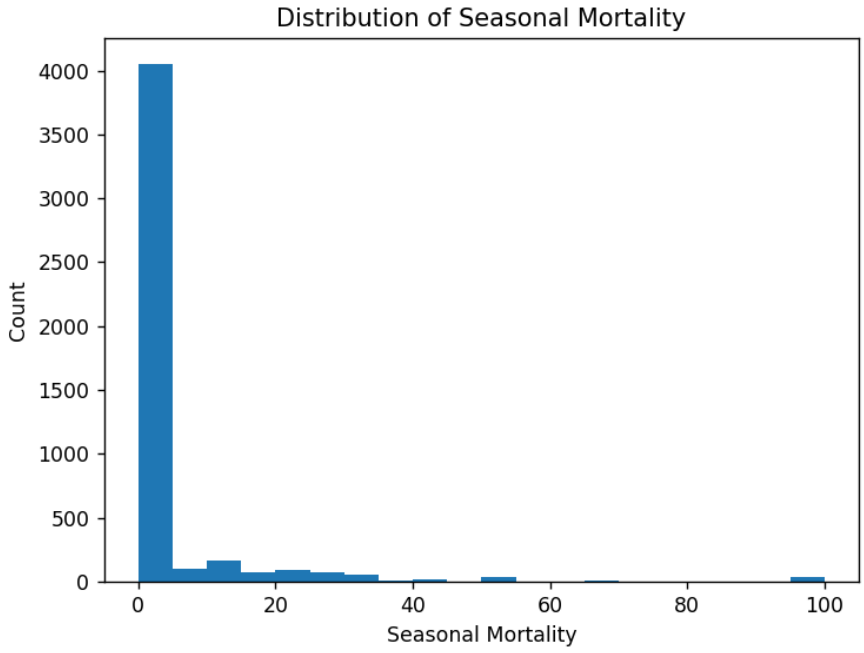

Figure 2. Distribution of existing honeybee hive mortality data showcasing a class imbalance.

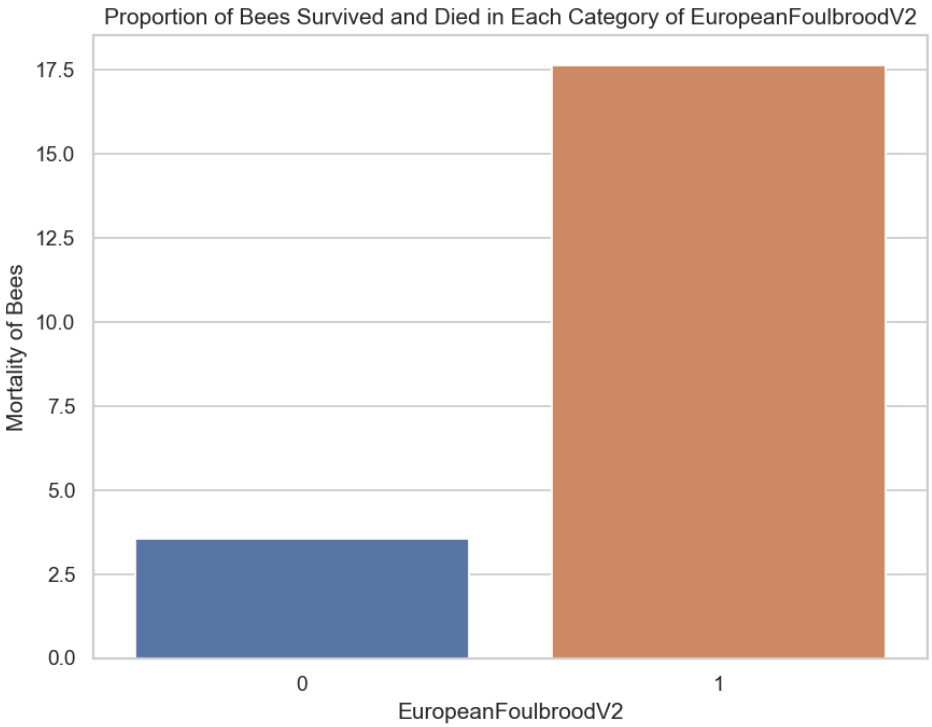

Figure 3. Initial data of beehives with existing European Foulbrood compared against the seasonal mortality of the hives.

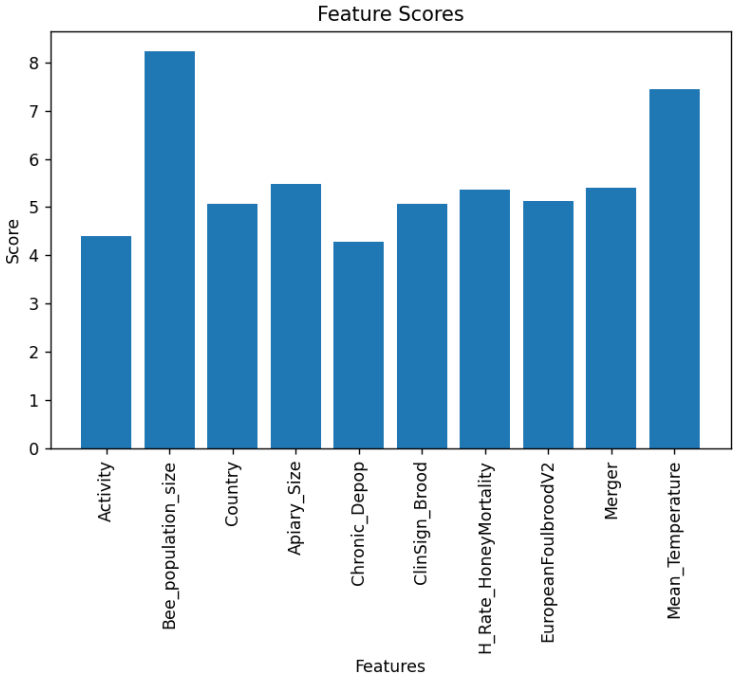

Figure 4. The top ten features in predicting seasonal mortality after undergoing feature selection.

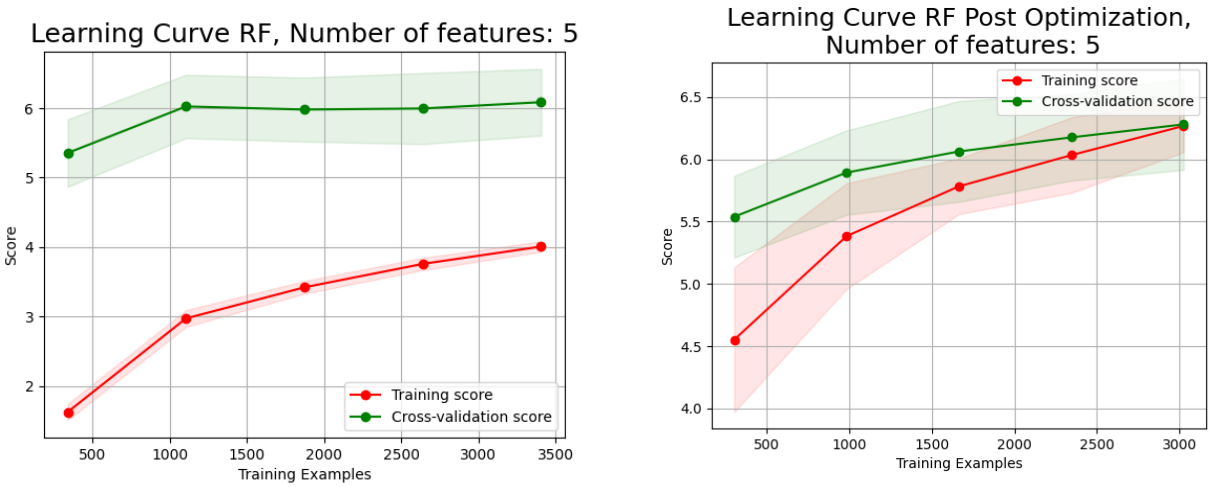

Figure 5. Before and after learning curves of random forest with 5 feature selection.

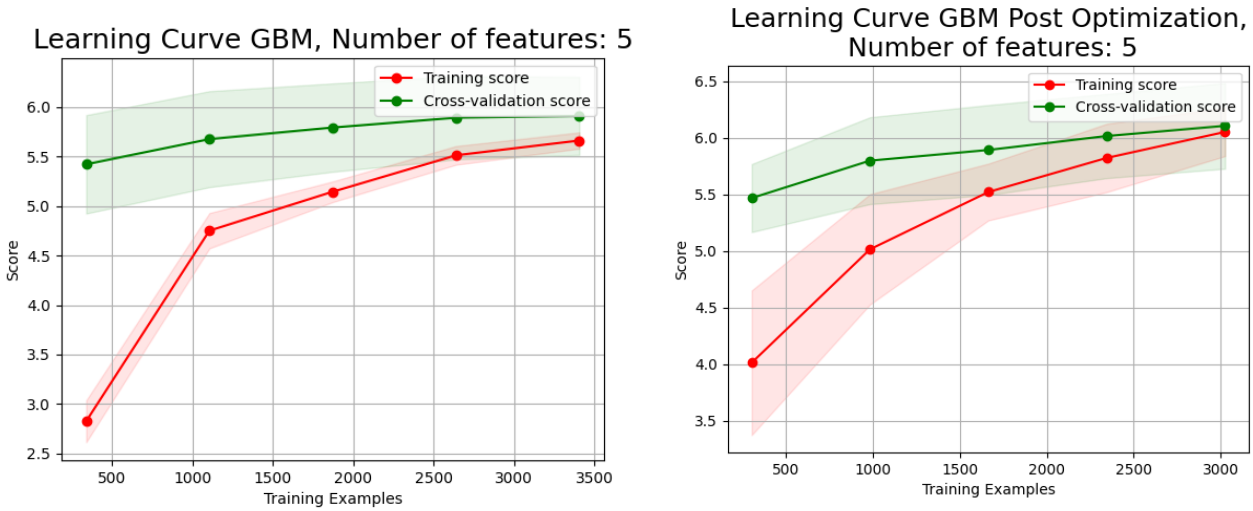

Figure 6. Before and after learning curves of gradient boosting machines with 5 features.

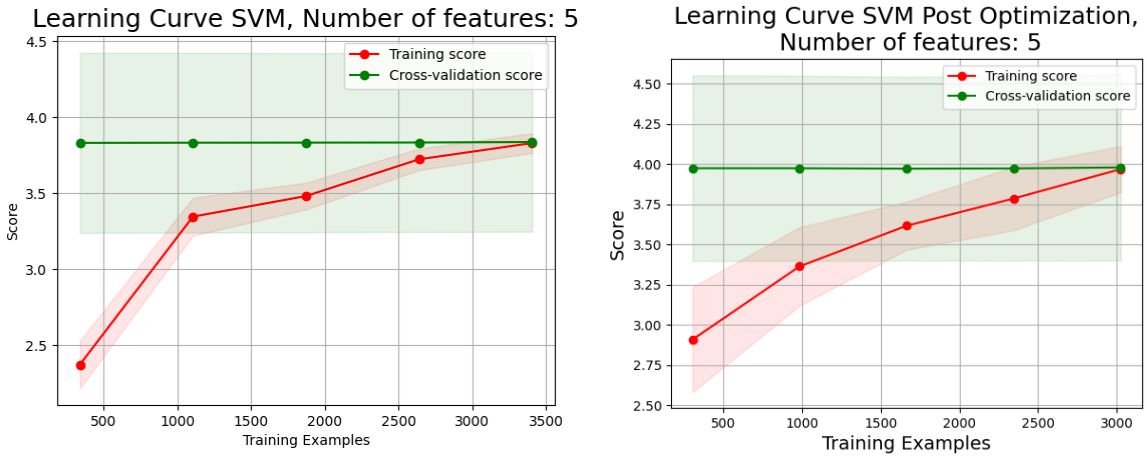

Figure 7. Before and after learning curves of support vector machines with 5 features.

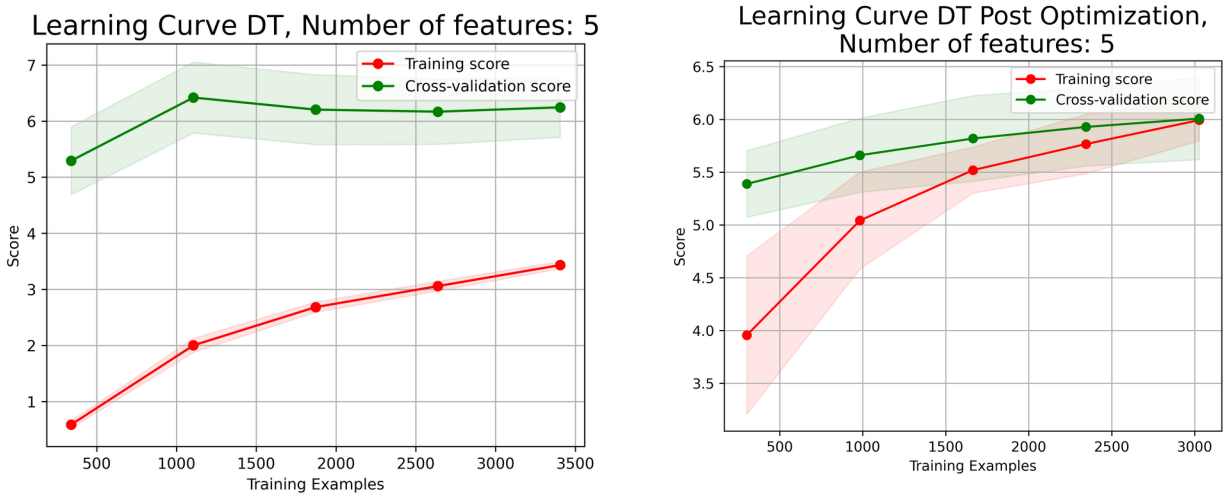

Figure 8. Before and after learning curves of decision trees with 5 features.

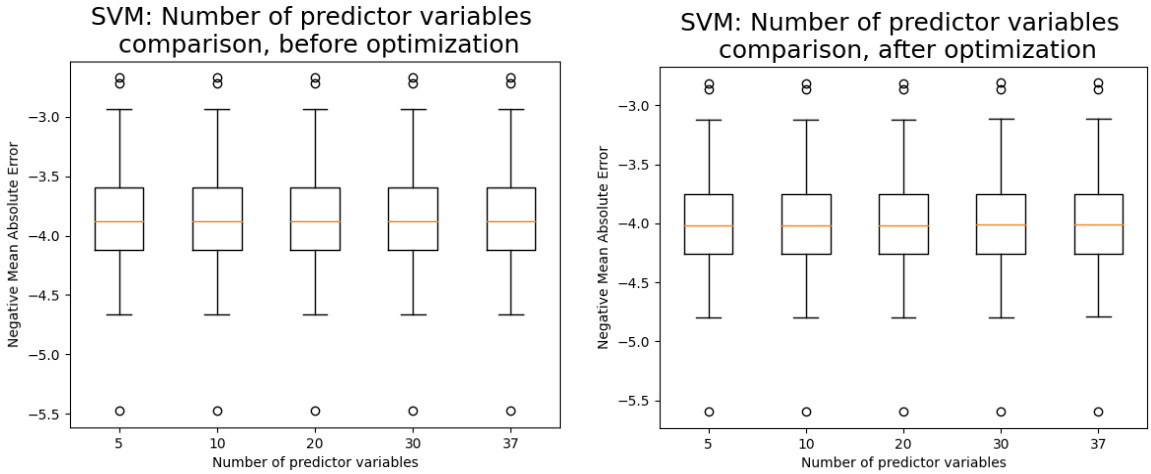

Figure 9. Boxplots of SVM negative mean absolute error training scores before and after hyperparameter optimization. Demonstrates the minimal effect that hyperparameter optimization had on SVM-based models.

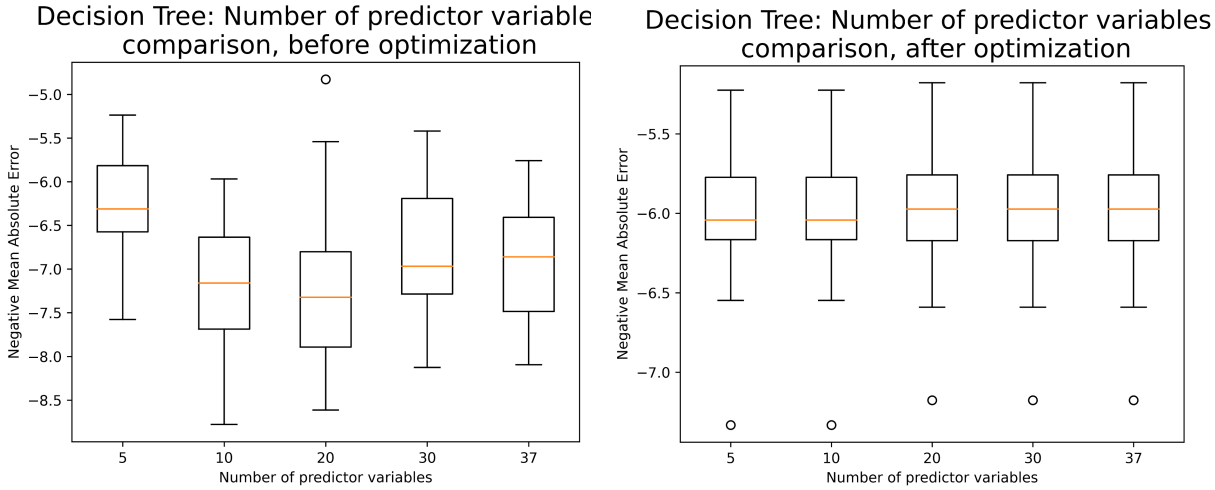

Figure 10. Boxplots of DT negative mean absolute error training scores before and after hyperparameter optimization. Shows the improvements to training score for each of the subsets.

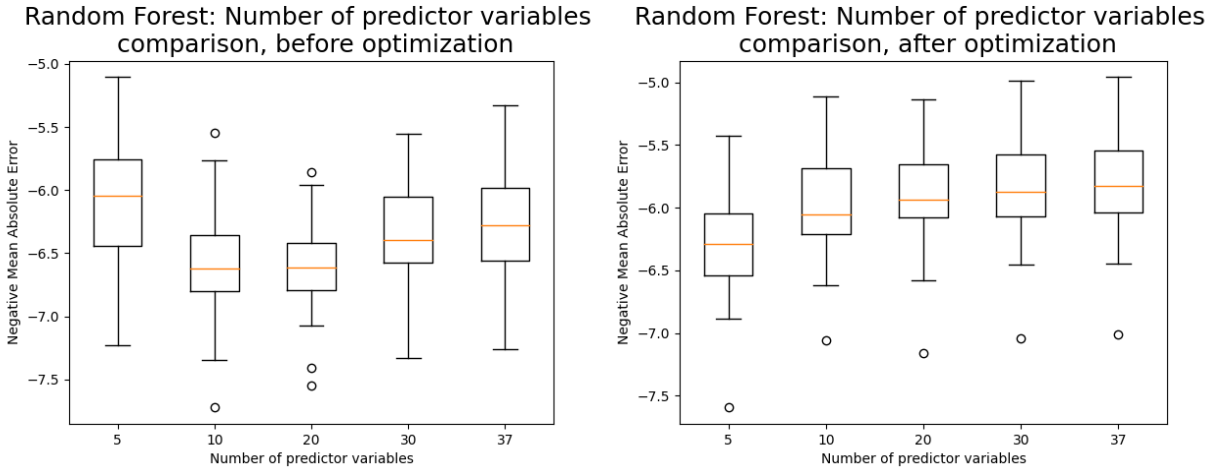

Figure 11. Boxplots of RF negative mean absolute error training scores before and after hyperparameter optimization. Shows the improvements to training score for each of the subsets with the exception of the subset with 5 features.

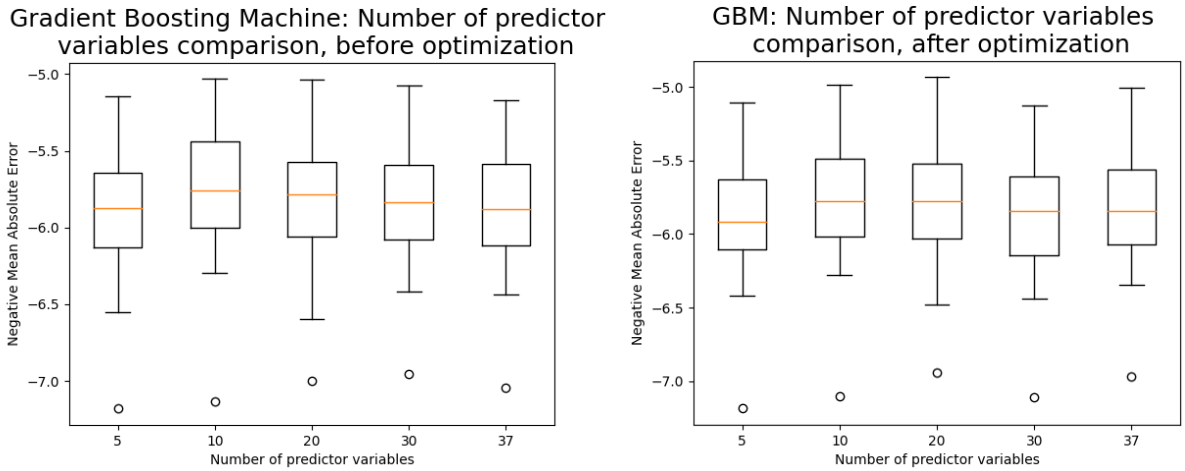

Figure 12. Boxplots of GBM negative mean absolute error training scores before and after hyperparameter optimization. Shows insignificant improvements after hyperparameter optimization.

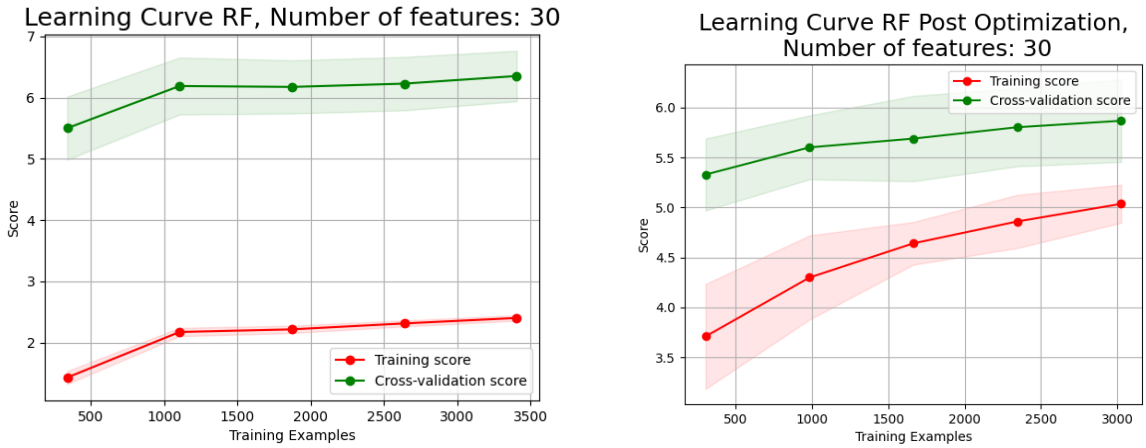

Figure 13. Before and after learning curves of random forest with 30 features. Demonstrates the underfitting that was apparent before and after optimization.

Tables

Table 1: Features in the EPILOBEE dataset

Beekeeper-related	**Description**
Age	Age of the beekeeper
Activity (Part-time, full-time, hobby)	Part-time, full-time, or hobby
Qualification	Formal qualification in beekeeping
Training	Attendance at beekeeping training in past 3 years

Coop-treat	Member of cooperative treatment against varroa
Bee population size	Total number of colonies owned by the beekeeper
Beekeep for	Experience beekeeping
Apiary size	Total number of colonies in the apiary
Apiarist book	Use of an apiarist book
Org member	Member of a beekeeping organization
Continue	Does the beekeeper plan to keep managing bees for more than two years
Management objective	Management goal of the beekeeper.
Bee and hive health-related	
Seasonal_mortality	The colony mortality during the beekeeping season
Breed	Breed of honeybee
Production	Products that are harvested from the hive
Chronic depopulation	Presence of chronic bee depopulation
Foulbrood	Signs of foulbrood

Clinical signs honeybees	Any clinical signs on the honeybees
High rate of colony mortality in the past year	Colony mortality greater than 10% in past year
High rate of honey bee mortality in the past year	Honey mortality greater than 10% in past year
Other event	Any other clinical event in the past year
Varroa mites	Presence of varroa mites in the past
Queen problems	Problems with queens
Swarms bought	Number of swarms purchased
Swarms produced	Number of swarms produced
Queen bought	Number of queens purchased
Queen produced	Number of queen produced
Varroosis	At least one colony suffering from varroosis
Chronic paralysis	At least one colony suffering from chronic paralysis
American foulbrood	At least one colony suffering from american foulbrood
Nosemosis	At least one colony suffering from nosemosis
European foulbrood	At least one colony suffereing from nosemosis

Merger	Have any colonies been merged together
Winter mortality	Colony mortality during the overwintering season
Geography-related	
Country	Which european country is the apiary located in
Environment	Environment surrounding the apiary
Migration	Apiary migrate at least once in the past year

Table 2: Hyperparameter optimization

Algorithm	Parameters	Hyperparameter tuning method
DT	Splitter: (best, random) max_depth: [1,3,5,7,9,11,12], min_samples_leaf: [0.00001, 1,2,3,4,5,6,7,8,9,10], min_weight_fraction_leaf : [0.00001, 0.1,0.2,0.3,0.4,0.5], min_leaf_nodes: [2,5,10,20,30,40,50,60,70,80,90]	GridSearchCV

GBM	learning_rate: [0.1, 0.05, 0.01], max_depth: [3, 5, 7], n_estimators': [100, 500, 1000], subsample: [0.8, 0.9, 1.0], min_samples_split: [2, 5, 10]	RandomizedSearchCV
RF	n_estimators: [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000], max_depth: [1, 11, 22, 33, 44, 55, 66, 77, 88, 99, 110, None], min_samples_split:[2, 5, 10], min_samples_leaf : [1, 2, 4], bootstrap : [True, False], max_features: ['sqrt', 'log2']	RandomizedSearchCV
SVM	C: [1, 5], gamma: [1, 5], epsilon: [0.1, 0.3], degree: [2, 4]}	RandomizedSearchCV

Table 3: MSE of each algorithm and subset size.

Algorithm	**Number of features:**				
	5	10	20	30	37
RF	111.0880200	115.5343047	109.5682821	106.1877027	106.3466581
DT	108.8336863	108.8336863	107.6217975	107.6217975	107.6217975
GBM	110.7894799	111.11170600	108.6028532	107.6944143	108.0053527
SVM	120.1525131	120.1536764	120.1521795	120.1623277	120.1608263