Dairy Cattle Artificial Insemination Probability Prediction Using Artificial Intelligence

Pratiksha Pradeep Sharma, Nikolay A. Bliznyuk, Albert De Vries

Agriculture and Biological Engineering Department, Gainesville, Florida, 32608

Abstract

Artificial insemination in the field of dairy science has been a boon for farmers, which allows them to breed the cows when they are in heat, as the heat detection is majorly done on the basis of the cow's increased activity, due to other biological factors of the cow there can be a lot of uncertainty in the cow having a positive breeding outcome. Hence, the time of the insemination matters a lot, which can be understood using data science methods. Specifically machine learning can be used to help solve this uncertainty issue. Machine learning is used to draw inferences from patterns in the data using various algorithms by incorporating the history, reproduction, and biological features of the cow the decision making for breeding and the issue of animal replacement can be solved, also while making the whole breeding process more economical for the farmers.

Data

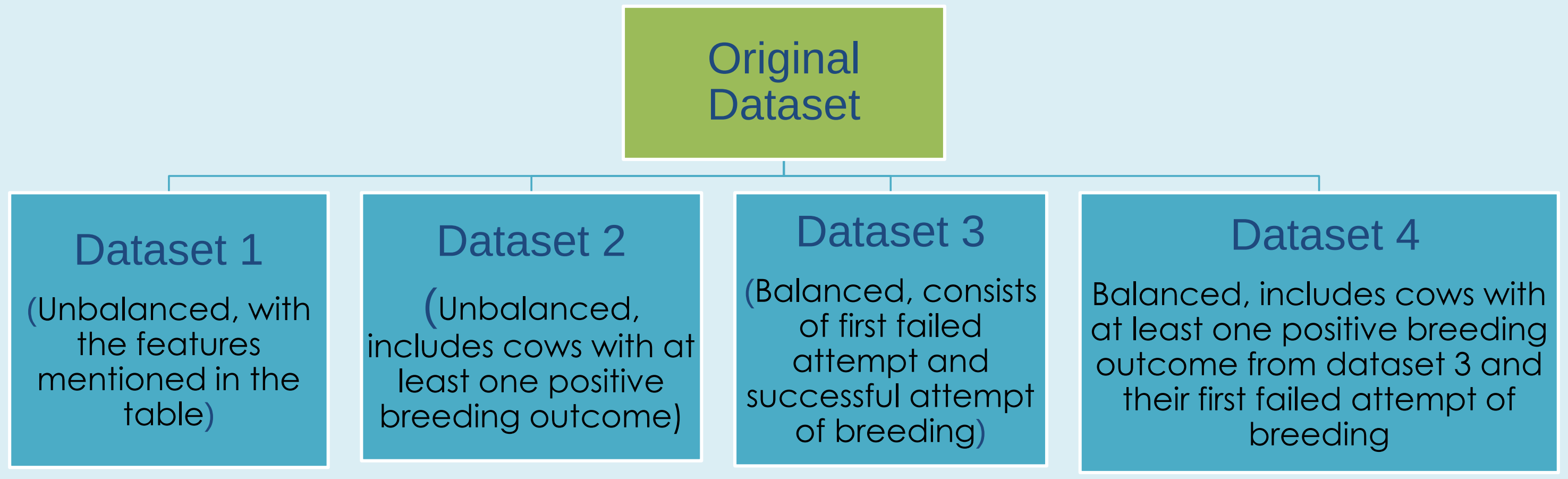
Data from cows is collected through the use of various types of sensors. The dataset available includes the activity, reproduction, health disease and milk yield information of the cows. This data contains 26,654 samples for 6632 cows. Current work is based on utilizing the activity and reproductive history of the cow. The activity related features of the cow include the steps, lying and standing time. The reproductive features are dim, nbred, and nlact, and the target is the breeding result the lagsteps contains missing and Nan values, and the dataset is highly imbalanced with the percentages of 0's being 78.8% and 1's being 21.1%.

Feature Engineering: The lagsteps feature contains values for the steps of the cow/hour and it is recorded for 10 days. The general expectation is that the steps should increase as the day of breeding approaches. A new feature called deviation is derived from the lagsteps which is the difference between the average steps/hour from day 2 to day 10 and readings from day 1, after which the breeding is performed. Season of the breeding, which is obtained from the date of the breeding.

Features	Detail
Lagsteps	Steps of the cow/hour
DIM, nbred, nlact, date	Days in milk, no. of breeding, no. of lactations, date of insemination

Features used for the models

Types of Datasets: Four datasets are used for comparing the results and understanding which models work the best for the research problem.



Original Dataset

Dataset 1 (Unbalanced, with the features mentioned in the table)

Dataset 2 (Unbalanced, includes cows with at least one positive breeding outcome)

Dataset 3 (Balanced, consists of first failed attempt and successful attempt of breeding)

Dataset 4 Balanced, includes cows with at least one positive breeding outcome from dataset 3 and their first failed attempt of breeding

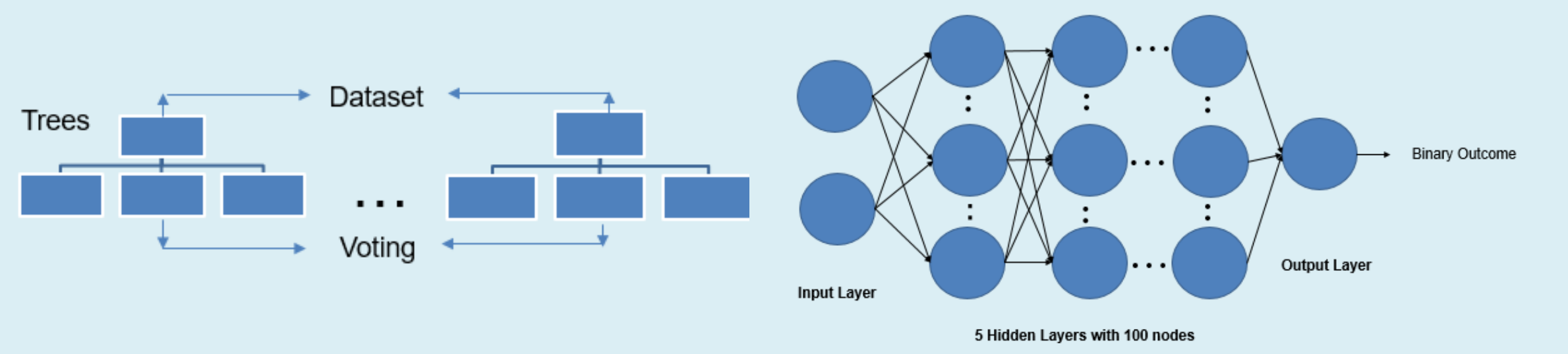
Methodology

Methods used for predicting the probabilities of the breeding outcome are:

1. **Logistic Regression** : It models the probability of distinct outcome given an input variable using the logistic function, and is therefore used for predictions of binary outcomes.
2. **Random Forest** : Comprises of countless individual decision trees, where each decision tree lets out a class prediction and the class with the most votes turns in to our model's expectation.
3. **XGBoost Classifier**: It is also a decision-tree based ensemble machine learning algorithm. It utilizes a gradient boosted framework. L1 and L2 regularization are used to improve the generalization capability of the model.
4. **Deep Neural Network** : The DNN consists of an input layer, hidden layers and an output layer. Gradient descent is used as an optimizer to update the weights.

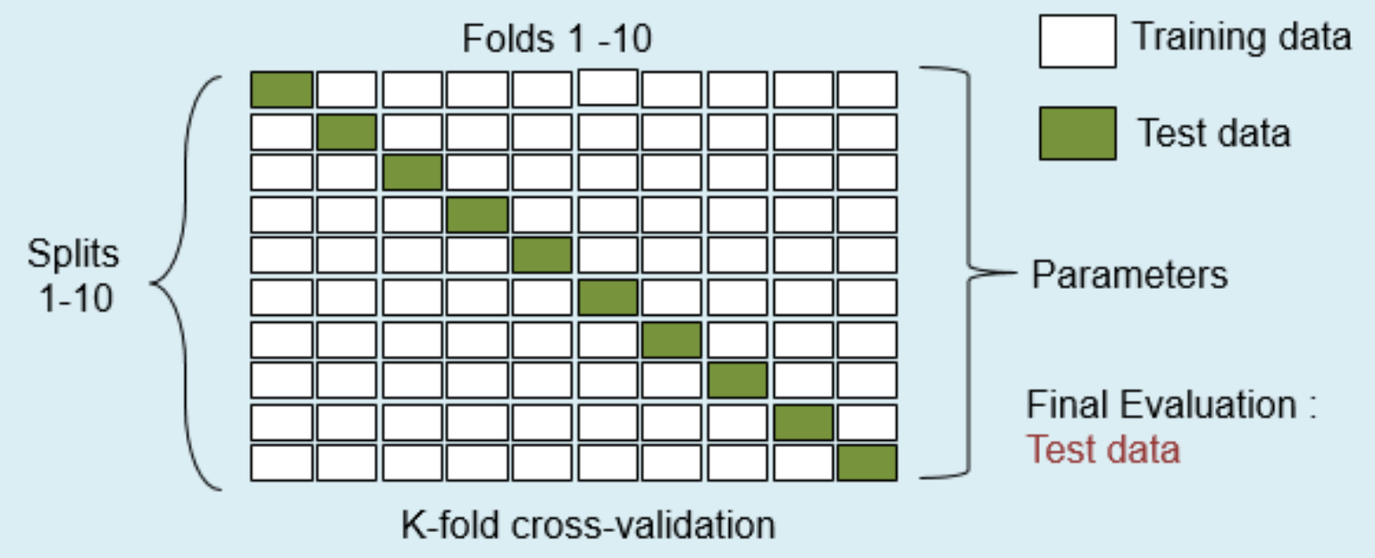
$$\text{Logistic}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

Logit Function



Random Forest

DNN



Method to avoid over-fitting

Evaluation Metrics: Metrics are parameters that can be utilized to check how accurately the particular problem at hand has been solved. The evaluation metrics used are ROC curve, AUC score, and confusion matrix.

Results

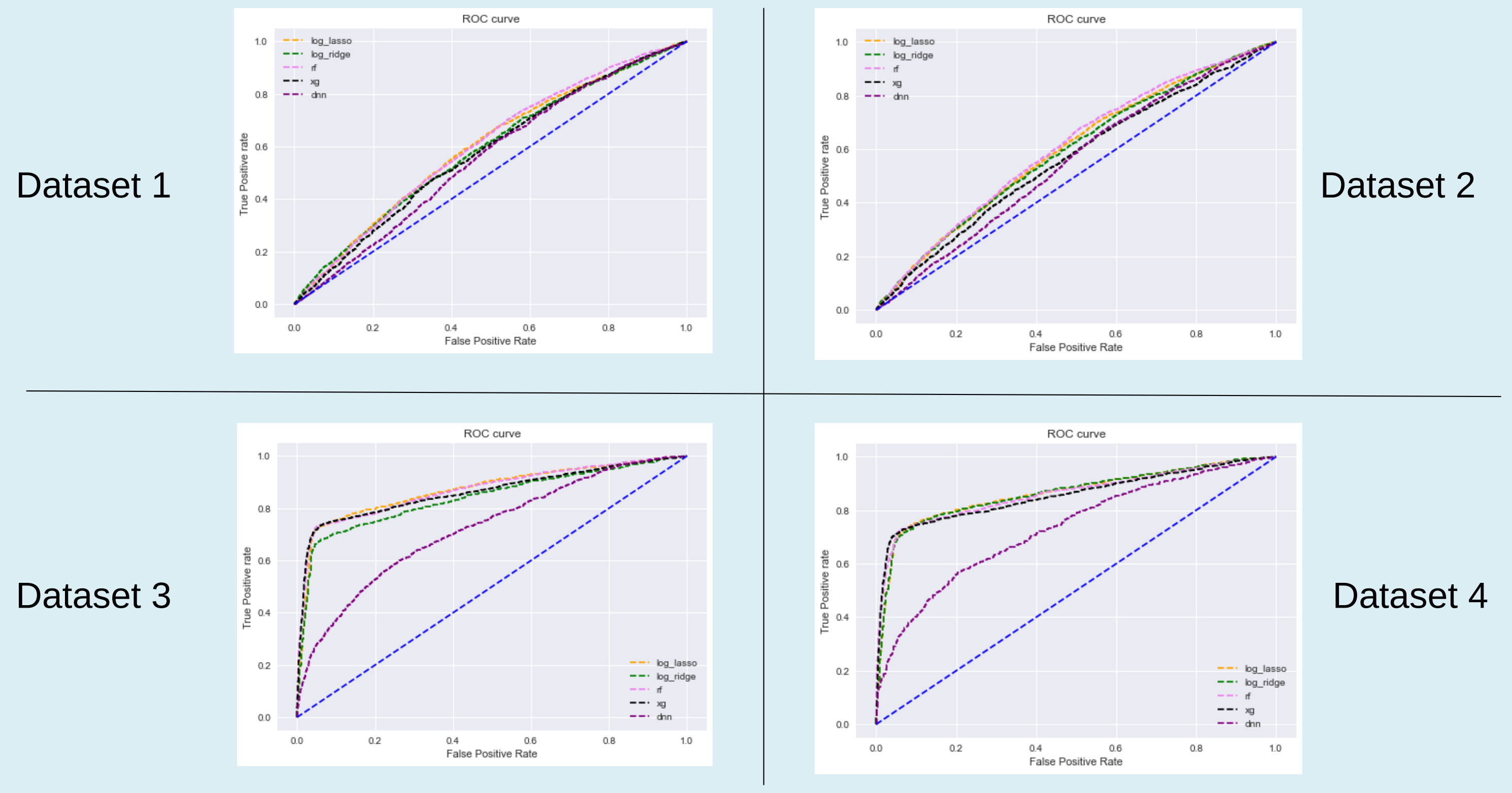
The results for four datasets obtained are:

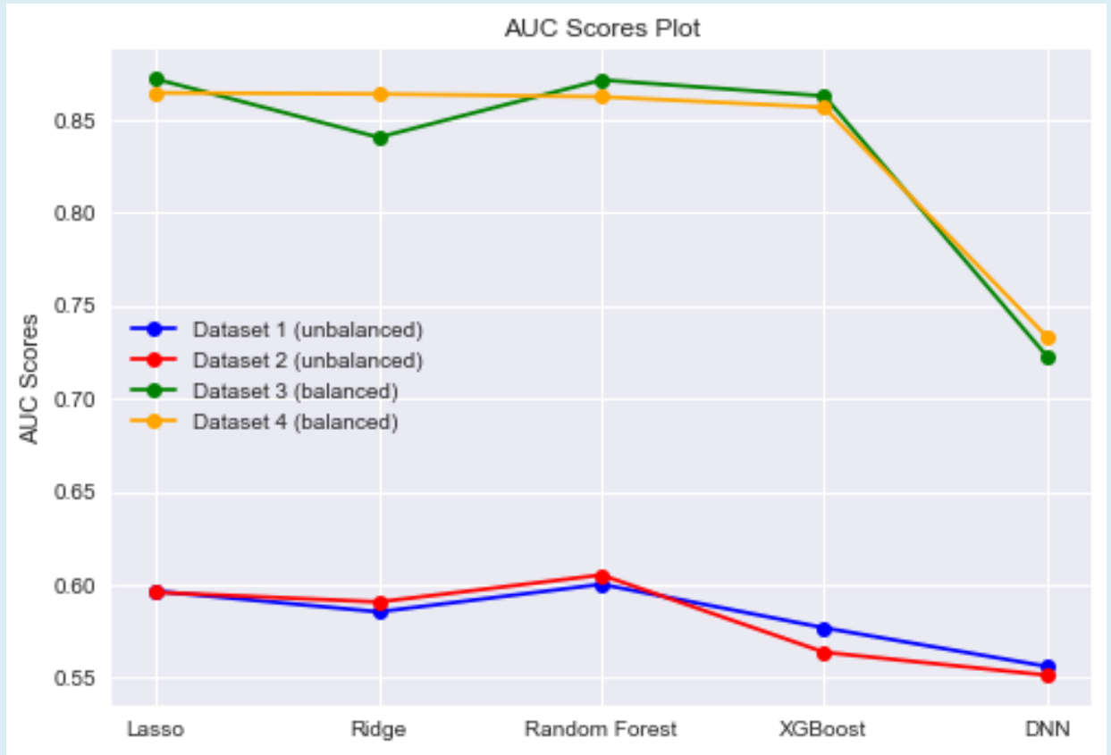
Predicted 0 1; Actual 0 1

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Logistic (Lasso)	6296 1725 1 1	4347 1667 32 29	1570 467 88 1128	1113 323 88 849
Logistic (Ridge)	6296 1725 1 1	4342 1661 37 35	1340 433 318 1262	1111 328 90 844
Random Forest	6297 1726 0 0	4373 1691 6 5	1559 449 99 1246	1121 318 80 854
XGBoost	6121 1650 176 76	3993 1461 386 235	1503 422 155 1273	1079 298 122 874
DNN	5650 1530 647 196	3569 1336 810 360	1051 548 607 1147	791 395 410 777

Results

ROC Curves :





Plot for AUC scores for all the datasets

Conclusion

The confusion matrix, ROC curves and AUC scores are obtained for four datasets, where two of them are balanced datasets.

1. For dataset 1 and 2 the classification results show that XGBoost and DNN give some class separation. But, the ROC curve and AUC scores show that Logistic regression with lasso penalty and random forest give better results, with the highest auc score being 0.6003 and 0.6053 respectively.
2. For dataset 3 and 4, which are the balanced datasets, random forest has the lowest FN, and FP after classification, while logistic regression with lasso penalty and random forest have the best AUC scores and ROC curves with highest auc scores being 0.8710 and 0.8620. In general the AUC scores for balanced datasets are much higher than the unbalanced datasets (dataset 1 and 2), with logistic regression (lasso penalty) and random forest being the most accurate while predicting the probabilities.

References

- Inchaisri C, De Vries A, Jorritsma R, Hogeveen H. Improved knowledge about conception rates influences the decision to stop insemination in dairy cows. Reprod Domest Anim. 2012 Oct;47(5):820-6. doi: 10.1111/j.1439-0531.2011.01975.x. Epub 2012 Jan 2. PMID: 22211392.
- Madureira AM, Silper BF, Burnett TA, Polsky L, Cruppe LH, Veira DM, Vasconcelos JL, Cerri RL. Factors affecting expression of estrus measured by activity monitors and conception risk of lactating dairy cows. J Dairy Sci. 2015 Oct;98(10):7003-14. doi: 10.3168/jds.2015-9672. Epub 2015 Aug 5. PMID: 26254517.
- Saleh Shahinfar, David Page, Jerry Guenther, Victor Cabrera, Paul Fricke, Kent Weigel,Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms, Journal of Dairy Science, Volume 97, Issue 2, 2014, Pages 731-742, ISSN 0022-0302

UF/IFAS AGRICULTURAL AND BIOLOGICAL ENGINEERING