**Using machine learning for imputation and predicting finisher performance in commercial pig production**

**Application:** The application of this research is to modernise and optimise pig production by harnessing data, technology, and advanced machine learning techniques for monitoring and prediction purposes. This enables farmers to efficiently allocate resources such as feed and labour while minimising wastage.

**Introduction:** The growing global population and climate challenges require livestock production systems to increase yields whilst also maintaining acceptable welfare standards. Therefore, accurately allocating resources such as feed and labour whilst minimising waste will be a top priority for livestock farmers. To achieve this, the collection and utilisation of high-quality quantitative data for monitoring and prediction purposes are imperative. Traditionally, predictive analytics in pig production have relied on general linear models to predict performance, using multiple predictor variables. However, these models often struggle to handle complex relationships between factors and non-linear data that often arise when multiple predictor variables are used (Comrie 1997). Recently, machine learning (ML) algorithms have emerged as an alternative to address these issues. Machine learning algorithms learn from historical data to make predictions, with their accuracy improving as more data is included. The aim of this study is to use ML algorithms on data collected from commercial farms to (1) address missing data issues, (2) obtain more accurate predictions compared to classic linear regression models and (3) identify the most important early-life predictor variables to predict finisher performance.

**Materials and methods**: Data were collected in four batches under commercial conditions at multiple sites of Karro Food Group Ltd. A total of 313 sows were balanced across parities 1 to 6 and were serviced with four different sire lines (A- Pietran synthetic terminal sire, B- Hampshire synthetic terminal sire, C-White synthetic terminal sire, and D-Hampshire synthetic terminal sire). A total of 5225 piglets were individually ear-tagged at birth. Individual birth weights (BWT) were collected from batches 2, 3 and 4, and individual weights were recorded for all batches at finishing entry (week 9; WK9) and finishing exit (week 23; WK23). Abattoir data including, net cold weight (NCW) (kg) and carcass lean meat (LM) (kg), were collected. Data underwent pre-processing, involving the encoding of categorical data, and outlier removal using the *z*-score algorithm. Batch 1 BWT was not measured, and data was missing completely at random for WK9, WK23, NCW and LM. The level of missing data gathered under commercial conditions in this study mirrors authentic on-farm conditions, thereby presenting a dynamic real-world context for predictive analysis. To address missing data, imputation was performed using ML algorithms from the Python scikit-learn library, including linear regression, Decision tree regression (DT Regression), k-Nearest Neighbour regression (KNNR) and Support Vector (SV) regression, Random Forest Regression (RFR), Extreme Gradient Boosting (XGBoost) regression, Light Gradient Boosting (LGBM) regression, Adaptive Boosting (AdaBoost) regression, Stacking Regression (SR) and Voting Regression (VR). A complete case dataset was randomly shuffled into 5 subsets were used to train and test the algorithms with the most accurate imputation algorithm selected using the relative root mean square prediction error and concordance correlation coefficient (You et al 2023). A final imputed dataset was made for downstream prediction analysis. The same ML algorithms used for imputation were used for finishing performance prediction. All model evaluations and statistical analyses were performed using Jupyter Notebook (Project Jupyter, USA). The accuracy of downstream prediction algorithms was assessed using mean absolute error (MAE), root mean squared error (RMSE), and the mean absolute percentage error (MAPE). The best predictor variables are listed in order of importance from highest to lowest.

Results

**Table 1.** Imputation and downstream analysis of BWT (required for finishing predictions) and finishing variables.

|  |  |
| --- | --- |
| *Imputation* | *Downstream analysis* |
| Variable | Missing data | Top imputer | Top predictive algorithm | RMSE(LR metric) | MAE(LR metric) | R2(LR metric) | Variable importances |
| BWT | 22.3% | SR\_SVR |  |  |  |  |  |
| WK9 | 13.8% | SR\_LR | SR\_SVR | 3.33(3.54) | 2.53(2.75) | 0.33(0.24) | imp\_BWT, PAR, SL, Batch, TLS, MED, Sex |
| WK23 | 23.6% | SR\_LR | SR\_LR | 9.45(10.1) | 7.09(7.70) | 0.29(0.17) | imp\_BWT, batch, SL, PAR, TLS, Sex, MED |
| NCW | 48.3% | RF | SVR | 6.66(7.25) | 4.85(5.57) | 0.29(0.17) | Rep, imp\_BWT, SL, Sex, PAR, TLS, MED |
| LM | 48.3% | RF | SR\_SVR | 1.84 (2.04) | 1.26(1.46) | 0.22(0.05) | SL, Sex, Batch, TLS, imp\_BWT, PAR, MED |

SR\_SVR = Stacking regression with SVR as the meta-model, SR\_LR = Stacking regression with LR as the metamodel, imp\_BWT = imputed BWT, SL = Sire line, TLS = Total litter size, PAR = Parity, MED = Medicated, LR accuracy metrics are presented in brackets for comparison with top performing model.

**Conclusions:** This study highlights the application of ML algorithms in imputing large datasets and predicting finisher performance. Our findings indicate that ensemble models such as stacking regression and random forest regression were the most accurate for imputation. The best algorithms for predicting finishing performance were stacking regression algorithms with LR and SVR as meta-models and SVR, outperforming other methods, including linear regression, by capturing underlying data patterns more efficiently. Our findings also reaffirm the critical role of BWT in influencing finishing performance outcomes.

Acknowledgements Supported in part by Innovate UK Knowledge transfer partnerships. Ethics approval; AWERB No. 012020MF

References [1] Comrie, A.C., 1997. Journal of the Air & Waste Management Association 47, 653-663. [2] You, K., Ellis, J.L., Adams, S., Sahar, M., Jacobs, M., and Tulpan, D. 2023. Animal, 100921