



## Data-Driven Classifiers for Predicting Grass Growth in Northern Ireland

McHugh, O., Liu, J., Browne, F., Jordan, P., & McConnell, D. (2020). Data-Driven Classifiers for Predicting Grass Growth in Northern Ireland: A Case Study. In M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, & R. R. Yager (Eds.), *IPMU: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems: Proceedings, Part I - 18th International Conference, IPMU 2020* (pp. 301-312). (Communications in Computer and Information Science; Vol. 1237 CCIS). Springer. Advance online publication. [https://doi.org/10.1007/978-3-030-50146-4\\_23](https://doi.org/10.1007/978-3-030-50146-4_23)

[Link to publication record in Ulster University Research Portal](#)

### Published in:

IPMU: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems

### Publication Status:

Published online: 05/06/2020

### DOI:

[10.1007/978-3-030-50146-4\\_23](https://doi.org/10.1007/978-3-030-50146-4_23)

### Document Version

Author Accepted version

### General rights

The copyright and moral rights to the output are retained by the output author(s), unless otherwise stated by the document licence.

Unless otherwise stated, users are permitted to download a copy of the output for personal study or non-commercial research and are permitted to freely distribute the URL of the output. They are not permitted to alter, reproduce, distribute or make any commercial use of the output without obtaining the permission of the author(s).

If the document is licenced under Creative Commons, the rights of users of the documents can be found at <https://creativecommons.org/share-your-work/licenses/>.

### Take down policy

The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact [pure-support@ulster.ac.uk](mailto:pure-support@ulster.ac.uk)

# Data-Driven Classifiers for Predicting Grass Growth in Northern Ireland: A Case Study

Orla McHugh<sup>1</sup>[0000-0002-7205-4767] and Jun Liu<sup>1</sup>[0000-0001-8859-5405] and Fiona Browne<sup>2</sup>[0000-0003-0443-4730] and Philip Jordan<sup>1</sup>[0000-0002-8973-4734] and Deborah McConnell<sup>3</sup>

<sup>1</sup> Ulster University, Northern Ireland, UK

<sup>2</sup> Datactics, Belfast, Northern Ireland, UK

<sup>3</sup> AFBI, Belfast, Northern Ireland, UK

**Abstract.** There are increasing pressures to combat climate change and improve sustainable land management. The agriculture industry is one of the most challenging areas for these changes, especially in Northern Ireland, as agriculture is one of the larger industries. Research has been carried out across the island of Ireland into methods of improving farm efficiency in multiple areas of farming, including livestock health, machinery improvements, and crop growth. Research has been carried out in this study into grass growth in the dairy farming sector, specifically within Northern Ireland. Grass growth prediction aims to inform farmers and policy makers in their decision-making process regarding sustainable land management in agriculture. The present work focuses on analysing and evaluating how data-driven classifiers can be used for grass growth prediction using the data related to soil content, weather, grass quality components *etc.* Four classifiers, namely Decision Trees, Random Forest, Naïve Bayes, and Neural Networks, are chosen for this purpose. Classification results based on a real-world data set are analysed and compared to evaluate and illustrate the performance and robustness of the classifiers. The results indicate that it is difficult to declare a single classifier with the highest performance and robustness. Nevertheless, it indicates that tree classification methods are better suited to the data to be studied, as opposed to probabilistic methods and weighted methods, *e.g.*, the naïve Bayes classifier obtained a predictive performance of 78% when classifying spring seasonal grass growth data.

**Keywords:** Climate Change, Grass Growth Prediction, Data-driven Classifier.

## 1 Introduction

Climate change is a global issue that has become more pressing in recent years. Countries around the world are adopting strategies to combat the effects of climate change and reduce their greenhouse gas (GHG) output. This includes the United Kingdom (UK) and Ireland, which currently have plans in place to reduce GHG emissions from the 1990 baseline [1]. However, these strategies have not been enough to significantly reduce the output, therefore, more needs to be done. Each region within the UK has specific targets to achieve including Northern Ireland (NI), which must

reduce its carbon emission by 35% by 2030, to meet UK targets [2]. In NI, one of the main contributors to GHG emissions is the agricultural sector, which produces almost 30% of the total NI output [2]. Dairy farming is one of the largest agricultural industries in NI. According to the Committee on Climate Change, agricultural emissions in NI have continuously increased since 2009 despite efforts to improve the efficiency in dairy farming [2]. Therefore, it is vital that tools and support are provided to farmers and stakeholders within the industry to inform them on solutions and actions that can improve farming efficiency and reduce emissions.

This study relates to the improvement of dairy farming efficiency by focusing on sustainable land management and examining grass growth which is one of the cheapest feed sources for livestock in NI [3]. Grass growth rates are variable across the year and depend on various factors, with some of the most influential factors being meteorological *e.g.*, rainfall, solar radiation, and temperature. NI has a temperate climate that allows for a long growing season. Soil conditions such as temperature and moisture also have an influence on grass growth, curtailing growth particularly when soils are oversaturated or excessively dry. Other factors relating to management also impact grass growth such as fertiliser application, grazing intensity, and grazing rotation length.

Grass related data have been collected by the Agri-Food and Biosciences Institute (AFBI) across NI in their GrassCheck project. AFBI is a research and development organisation that supports the Department for Agriculture, Environment, and Rural Affairs (DAERA) and other UK government bodies and public organisations. The GrassCheck project consists of farmer research gathered across 50 locations in NI including beef, sheep, dairy, and crop plot farming. The project will run for three years from 2018 to 2020 collecting grass growth data, grass quality data, grazing event data, and meteorological data.

The authors of this research have performed an exploratory statistical analysis of the GrassCheck dataset detailed in [4]. In this study, the R programming language was used to provide a statistical overview, correlation analysis, and linear regression analysis of the GrassCheck data to identify the grass growth predictive features. A boxplot visualising the variance in the grass growth features (including pre-grazing cover, utilisation, and soil moisture) illustrated the variability of grass growth over an 8-month period in which data was recorded. A correlation analysis identified strong positive relations between offtake, pre-grazing cover and grass growth and strong negative relations between post-grazing cover and grass growth. Linear regression was performed on the GrassCheck dataset to determine which features had the greatest influence on grass growth. Using this method, pre-grazing cover and the available amount of grass to livestock (known as available) features were shown to be the best fit models when used as the explanatory variables. Other statistically significant features include offtake, utilisation, and month [4]. However, this study is still limited in finding the in-depth pattern for grass growth prediction. Advanced data analytics are expected to further enhance predictions by using, for example, data-driven classification models such as neural networks, naïve Bayes, and decision trees to expand on the exploratory statistics used to analyse grass growth data.

Therefore, the aim of this research was to aid in understanding how various grassland features contribute to the prediction of grass growth, and to analyse and evaluate various classification models to deduce which are the most suitable for grass growth prediction.

This paper is organised as follows; Section 2 provides an overview of related research in the area. Section 3 provides the methodology underpinning the research, with results and discussion presented in Section 4. Conclusions and future work are discussed in Section 5.

## 2 Related Work

In Ireland, research has developed a grass growth prediction model for dairy based farming [5] [3]. The Moorepark St. Gilles Grass Growth Model, known as the MoSt GG model, is a descriptive model providing insight into grass growth at paddock levels in Ireland. There are various inputs into this model including forecasted meteorological data, management strategy information, and fertiliser application, specifically nitrogen ( $N$ ). The outputs of this model include daily grass growth,  $N$  information such as the soil content, grass content, grass uptake, and nitrate leaching. The output from the model was compared to the output from an experimental farm in Cork, Ireland, for a period of two years. It was observed that, while the model was successful in improving some areas of prediction from a previous model, *i.e.*, better prediction of production per cutting date and per plot, it was not always accurate in others. For example, the  $N$  prediction in grass content and nitrate leaching was underestimated, potentially since the model does not consider previous years management techniques. Although this model was not designed for NI, the same principles can be applied to aid constructing a decision support system to support sustainability in NI. The decision support system could be expanded to make predictions to support farmers and policy makers in their decisions regarding sustainability.

Classification approaches, such as decision trees, artificial neural networks, and support vector machines have been used in multiple research studies for different classification problems. These include agricultural issues such as crop disease prediction [6], crop yield prediction [7], and grassland biomass estimation [8]. This research discusses the various classification methods used in agricultural prediction. When predicting crop disease, multiple classification methods were used including, neural networks, naïve Bayesian, random forest, decision trees, support vector machines, k-nearest neighbor, and ensemble models [6]. In this study, it was found that random forest and Gaussian naïve Bayes classifiers performed better than other classifiers when predicting binary data, while neural networks and random forest were better when predicting the original dataset. Multiple linear regression and density-based clustering classification methods have been used in this research area [7]. Multiple linear regression, neural networks, and adaptive neuro-fuzzy inference systems were also used in the area of grassland biomass estimation [8]. This research highlighted the use of the neuro-fuzzy system as it performed better when estimating biomass than the artificial neural networks and the multiple linear regression.

### 3 Materials and Methods

#### 3.1 Data

The data used in this research is grassland data provided by AFBI, from the GrassCheck project. Different features have been collected including grass growth, grass quality, grazing events, and meteorological data. The features within this dataset have been outlined in Table 1 below.

**Table 1.** List of features in the datasets

Feature	Description	Type	Unit	Example
ID	The farm identifier assigned to the farms	Numeric		1-12, 41-48
Month	The month that the grass growth took place	Numeric		3 (March)
Week	The week that the grass growth took place	Numeric		12
County	The county in NI in which the farm is located	Categorical		Down
Grass Growth	The daily grass growth averaged across all paddocks on the farm	Numeric	Kg DM/Ha	Range: 0 – 145.3
Field	Farmer description of field where grass quality measurements are recorded	Qualitative		“Yard Field”
Conditions	Farmer description of weather conditions on day of measurement	Qualitative		“Bright and Sunny”
Dry Matter (DM)	The proportion of total grass components (fibers, proteins <i>etc.</i> ) remaining in the grass after water is removed	Numeric	%	Range: 9.6 – 40.4
Crude Protein	The protein content in the grass, minus effluent losses	Numeric	%	Range: 9.7 – 26.8
Acid Detergent Fiber (ADF)	Measurement of digestibility, via the cellulose, lignin, and lignified nitrogen content of the grass	Numeric	%	Range: 2.6 - 37
WSC	The soluble sugars released from the grass in the animal	Numeric	%	Range: 0 – 23.5
Metabolisable Energy (ME)	The energy content of the grass measured in megajoules of energy per kilogram of dry matter (MJ/Kg/DM)	Numeric	MJ/Kg/DM	Range: 9.9 – 13.4

Paddock	Farmer description of paddock where grazing event took place	Qualitative			“MC2B”
Event	Type of grazing event <i>i.e.</i> , grazed, or part grazed by animals or cut by machinery	Quantitative			Grazed
Date	The date of the grazing event	Date			30/03/18
Notes	Any notes made by farmer relevant to the event	Qualitative			“part grazed – value changed”
Pre-Grazing Cover	The amount of grass in the paddock immediately before animal grazing	Numeric	Kg DM/Ha		Range: -2200 - 8100
Post-Grazing Cover	The amount of grass in the paddock after animals have grazed	Numeric	Kg DM/Ha		Range: 50 - 4920
Available	The amount of grass available to livestock in the paddock, calculated as pre-grazing cover – 1500	Numeric	Kg DM/Ha		Range: -3700 - 6600
Offtake	The amount of grass removed by animals at that grazing event, calculated as the pre-grazing – post-grazing cover	Numeric	Kg DM/Ha		Range: -4012.5 - 6889
Utilisation	The amount of grass consumed by animals ( <i>i.e.</i> , the offtake) as a proportion of that available	Numeric	Kg DM/Ha		Range: -174 – 6.33
Total Rainfall	The total rainfall fallen on the day of grazing	Numeric	mm		Range: 0 – 36.4
Air Temperature	The average air temperature on the day of grazing	Numeric	°C		Range: -2.03 – 25.33
Solar Radiation	The average solar radiation on the day of grazing	Numeric	W/m <sup>2</sup>		Range: 0 - 610.375
Soil Moisture	The average moisture levels in the soil on the day of grazing	Numeric	cb		Range: -1.09 - 200

Note: Kg DM/Ha : Kilogram Dry Matter per Hectare; WSC: Water Soluble Carbohydrate

There is a total of 4917 records that have been labelled using the classifications of High, Medium, and Low. The data have been binned into these labels based on calculating interquartile ranges on the grass growth value, where the lower quartile is 32.5 Kg DM/Ha and the upper quartile is 75.9 Kg DM/Ha. Approximately 25% of the records are low, 50% are medium, and 25% are high. In numerical terms, this equates to 1146 low records, 2336 medium records, and 1153 high records. The prediction of grass growth can help farmers make management decisions about grazing, cutting, and other areas of farming decisions in order to improve farm efficiency. For instance, knowing that there will be a low grass growth rate in the next month can allow farmers

to ensure they have adequate stocks of feed concentrates to ensure the wellbeing of their livestock. There are some grass growth entries that are missing, which results in 282 records with an unknown classification category. These unknown variables were removed from analysis during this study. The missing variables have been introduced through the method of data collection used to collect the data which relied on individual farmer input. Missing data was also introduced via the grass growth dataset being measured daily, while the grass quality and grazing events were not measured daily, but measured more sporadically. This meant when the datasets were joined on the Farm ID and the Date, there were empty variables where there was no recordings in the grass quality and grazing event datasets.

### 3.2 Data-driven classifiers overview

The following prediction models were chosen due to their ease of use, and popularity within the predictive analytics domain and application in the agricultural industry.

**Decision Tree (DT).** DT classifies instances via a tree structure, where individual attributes are represented by nodes, and there are links between nodes. The DT calculates the information gained from the attribute and makes decisions based on which attribute has the most information gain [6].

**Random Forest (RF).** The RF can be described as a collection of individual decision trees that work together as an ensemble [6]. This classification model is useful as each individual decision tree is unlikely to make the same mistakes as the others, and therefore, the classification is safer from error.

**Naïve Bayes (NB).** NB is a probabilistic classifier that assumes attributes are independent of each other and they carry the same weight when making predictions [9].

**Neural Network (NN).** A NN is a classifier that, like a decision tree, uses nodes and links to make predictions. However, each node is assigned a weight, with priorities at each node split being given to the feature that has a larger weight [9].

### 3.3 Case studies for grass growth prediction

Two case studies were carried out in the analysis of the GrassCheck dataset. Firstly, analysis of the 2018 dataset was performed where missing data were included (4917 instances, 19 features). Secondly, the same dataset was analysed where instances containing missing data were removed (107 instances, 19 features).

The datasets were divided into seasonal data (winter data were excluded as there is no grass growth during these months and, therefore, no recordings take place). The dataset was divided into Spring (March - May), Summer (June - August), and Autumn (September - October). This resulted in imbalanced datasets as there is more likely to be high growth rates in summer months and lower growth rates towards the cooler times of year. To resolve this imbalance, the larger sets could have been reduced to the approximate size of the smallest set, resulting in even divisions. This method was not applied as it would result in a dataset that is too small to perform classification models. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) [10] was applied to the smallest set in the dataset in order to make synthetic data that resembled actual

data in the dataset. In the Spring dataset, SMOTE was applied at 250%, which means the smallest dataset is increased by 250%. SMOTE was applied again at 35%, resulting in approximately balanced numbers of Low, Medium, and High labelled data (384, 376, 273, respectively). In the Summer dataset, SMOTE was applied at 150%, and 90%, to result in approximately balanced numbers of Low, Medium, and High data (642, 675, 665, respectively). In the Autumn dataset, SMOTE was applied at 100%, 250%, and 60%, to result in approximately balanced numbers of Low, Medium, and High data (299, 292, 291, respectively).

For each case study, ten-fold cross validation was carried out to evaluate each classifier. The evaluation metrics in this research are Kappa statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Precision, Recall, F-Measure, and ROC Area. Several features in the dataset were determined to not have informational values, including Farm ID, Field, Conditions, Paddock, and Notes. As discussed in the statistical analysis performed on the data [4], some features have a stronger correlation with grass growth. This was further analysed using feature selection methods including the Pearson's Correlation Coefficient, Information Gain Attribute Evaluation, and Wrapper Subset Evaluation.

## 4 Results and Discussions

This section summarises the results and discusses the outcome of the experiments. The analysis performed consists of four classifiers used on multiple variations of the dataset. This includes yearly divisions and period divisions of spring, summer and autumn. Classification analysis was performed on the data using techniques including DT, RF, NN, and NB. The tables below displays the outcome of the analysis, by showing the percentage of correct predictions, Kappa statistics, MAE, RMSE, precision (P), recall, F-measure (FM), and ROC, respectively.

### 4.1 Yearly Analysis

The tables below (Table 2 and Table 3) display information from the classification of data over the year of 2018. Table 2 shows the evaluation metrics on the whole dataset without handling the missing data, which contains 4917 instances. Table 3 displays the evaluation metrics on the same dataset, but with all instances including the missing variable removed, resulting in classification being performed on 107 instances.

**Table 2.** Evaluation metrics on the whole dataset.

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	<b>73</b>	<b>0.5624</b>	<b>0.2511</b>	0.3567	0.737	<b>0.735</b>	<b>0.731</b>	0.845
RF	<b>73</b>	0.546	0.2784	<b>0.3523</b>	<b>0.753</b>	0.734	0.725	<b>0.871</b>
NB	61	0.3497	0.2998	0.4306	0.628	0.610	0.599	0.738
NN	59	0.3303	0.2995	0.4386	0.613	0.599	0.596	0.731



**Table 3.** Evaluation metrics on the dataset excluding the variables with the missing data.

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	65	0.405	0.2914	0.4188	0.678	0.654	0.631	0.695
RF	80	<b>0.6676</b>	0.2376	<b>0.3211</b>	<b>0.809</b>	<b>0.804</b>	<b>0.794</b>	<b>0.917</b>
NB	62	0.3816	0.2504	0.4385	0.622	0.617	0.619	0.772
NN	76	0.5934	<b>0.1732</b>	0.3753	0.757	0.757	0.752	0.863

Table 2 shows that the tree classifiers, *i.e.*, DT and RF, have the best metrics out of the four classifiers. DT has the greatest Kappa statistic of 0.56, lowest MAE of 0.25, and highest recall and F-measure of 0.74 and 0.73, respectively. RF shows the best performance in RMSE with 0.35, precision with 0.75, and ROC area with 0.87. The results show that, overall, the RF classification method performed the best over the two datasets as the ROC curve is the highest in both sets at 0.871 and 0.917. It also has the lowest RSME across the classifiers at 0.3523 and 0.3211. However, RF is susceptible to influence by an imbalanced dataset, *i.e.*, the Medium category is a larger set than the other two growth rates, resulting in a skewed output.

The NN classifier showed the greatest improvement when missing data were removed from the dataset in terms of all metrics, *e.g.*, it increased from 0.33 to 0.59 in Kappa statistics, while reducing the MAE from 0.30 to 0.17, from Table 2 to Table 3, as a networks performance will increase when all features are available, *i.e.*, when there are no missing data. The NB classifier showed little difference in terms of performance when comparing the analysis using the full dataset, to the dataset where missing values are with no strong improvement observed. This is due to the assumption of independence of the attributes, as not all the attributes in these data are independent, *e.g.*, offtake, available, and utilisation depend on the pre-grazing and post-grazing cover.

## 4.2 Seasonal analysis

The tables below (Table 4, Table 6, and Table 8) display information from the classification of data pertaining to spring, summer, and winter. Table 5, Table 7, and Table 9 show the evaluation metrics on the spring, summer, and winter dataset, respectively, where SMOTE has been applied to balance the classes in the datasets, resulting in balanced categories of low, medium, and high.

**Table 4.** Evaluation metrics for Spring dataset (normal).

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	79	<b>0.6468</b>	0.1944	<b>0.3252</b>	<b>0.794</b>	<b>0.790</b>	<b>0.790</b>	0.874
RF	79	0.637	0.2268	0.3268	0.787	0.786	0.786	<b>0.888</b>
NB	63	0.3743	0.2458	0.4393	0.635	0.627	0.616	0.801
NN	78	0.6304	<b>0.1565</b>	0.3547	0.782	0.782	0.782	0.865

**Table 5.** Evaluation metrics for Spring dataset (SMOTE).

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	84	0.7594	0.1523	0.2889	0.842	0.840	0.839	0.921
RF	85	<b>0.7726</b>	0.1519	<b>0.2653</b>	<b>0.850</b>	<b>0.848</b>	<b>0.849</b>	<b>0.958</b>
NB	78	0.666	0.1605	0.3399	0.784	0.777	0.777	0.915
NN	84	0.7567	<b>0.1154</b>	0.3058	0.838	0.838	0.838	0.934

All the classifiers were improved from Table 4 to Table 5, in terms of correctly predicting instances when the dataset classes have been balanced, *e.g.*, NB showed the largest increase of 15%, from 63 to 78%. However, this does not mean that NB is a good classification method for these data. Although it has shown the most improvement in all features, including MAE and RMSE (reduction of 0.09 and 0.10 respectively), it is the overall least successful when predicting the level of grass growth, again due to the assumption made of independence. RF could be considered a good performer as second to NB, as it improved the most across most of the metrics. For instance, the Kappa statistic increased by 0.14, and the precision and recall have increased by 0.062 and 0.063, respectively. Overall, there are minor differences between DT, RF, and NN as they have similar evaluation outputs across all of the metrics.

**Table 6.** Evaluation metrics for Summer dataset (normal).

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	77	<b>0.6047</b>	0.2112	0.345	<b>0.769</b>	<b>0.768</b>	<b>0.764</b>	0.847
RF	75	0.5655	0.236	<b>0.3403</b>	0.754	0.748	0.741	<b>0.868</b>
NB	64	0.4402	0.2686	0.4153	0.671	0.638	0.640	0.797
NN	74	0.5496	<b>0.1967</b>	0.3943	0.739	0.737	0.731	0.810

**Table 7.** Evaluation metrics for Summer dataset (SMOTE).

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	80	0.6926	0.1622	0.3252	0.798	0.795	0.796	0.907
RF	84	<b>0.7607</b>	0.1619	<b>0.2768</b>	<b>0.848</b>	<b>0.841</b>	<b>0.841</b>	<b>0.952</b>
NB	71	0.572	0.1949	0.3655	0.721	0.714	0.702	0.903
NN	82	0.7365	<b>0.1242</b>	0.321	0.829	0.824	0.825	0.923

In Table 6, DT has the largest percentage of correctly predicted instances of 77%. It also has the best performance in Kappa statistic, precision, recall, and F-measure (0.60, 0.77, 0.77, 0.76, respectively). However, when SMOTE is applied in Table 7 to balance the classes in this dataset, RF becomes the better classification method as it has the best metric value in Kappa, RMSE, precision, recall, F-measure, and ROC area, (0.76, 0.28, 0.85, 0.84, 0.84, and 0.95 respectively). RF shows the greatest improvement in six of the eight metrics, including precision (0.75 to 0.85), recall (0.75 to 0.84), and F-measure (0.74 to 0.84). This means the classifier is returned more accurate results. NN shows

the greatest improvement in RMSE, where it reduced from 0.39 to 0.32, and in the ROC area, where it increased from 0.91 to 0.92. Overall, all of the classification methods have improved in all of the evaluation metrics with the addition of synthetic data.

**Table 8.** Evaluation metrics for Autumn dataset (normal).

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	80	<b>0.627</b>	0.1861	0.3296	<b>0.799</b>	<b>0.804</b>	<b>0.794</b>	0.834
RF	78	0.5827	0.2111	<b>0.323</b>	0.783	0.780	0.776	<b>0.875</b>
NB	48	0.2083	0.3471	0.5412	0.665	0.483	0.533	0.693
NN	75	0.5382	<b>0.1685</b>	0.369	0.749	0.752	0.750	0.846

**Table 9.** Evaluation metrics for on Autumn dataset (SMOTE).

Class	Acc (%)	Kappa	MAE	RMSE	P	Recall	FM	ROC
DT	84	0.7534	0.1513	0.3043	0.836	0.836	0.836	0.904
RF	84	<b>0.7567</b>	0.1726	<b>0.2822</b>	<b>0.839</b>	<b>0.838</b>	<b>0.838</b>	<b>0.946</b>
NB	67	0.5129	0.2185	0.4351	0.712	0.675	0.664	0.874
NN	81	0.7177	<b>0.1307</b>	0.3233	0.810	0.812	0.810	0.918

Table 8 shows the evaluation metrics of the classifiers before SMOTE has been applied, and indicates that the DT classifier could be considered the most appropriate method due to its performance across the metrics. For example, it has the greatest Kappa statistic of 0.63, the highest precision and recall rates of 0.80 each, and F-measure which is 0.79. However, when the synthetic data were added to the dataset, RF became the most accurate classifier due to its performance in Kappa, RMSE, precision, recall, F-measure, and ROC area (0.76, 0.28, 0.84, 0.84, 0.84, and 0.95 respectively). Again, NB showed the greatest improvement out of the four classifiers, with improvements across all of the features including Kappa statistic (increase of 0.30), MAE (reduction of 0.13), and RMSE (reduction of 0.11). However, NB was the least accurate classification method for this dataset as DT, RF, and NB, performed better across all of the metrics, *e.g.*, NB has an MAE of 0.22 when SMOTE was applied, while the other classifiers have an MAE of 0.17 or below. NN showed the greatest improvement in precision as it increased from 0.67 to 0.71, although the increases in each of the classifiers in this metric were very minor.

### 4.3 Further study and discussions

Three methods of feature selection were carried out as well on the dataset with no missing values including the Correlation Attribute Evaluation. This method is also known as the Pearson's Correlation Coefficient, in which attributes are ranked on how much information they provide to the prediction of the target class. The results of this

method show there is more of a correlation between Dry Matter, Soil Moisture, Offtake, Pre-Grazing Cover, Available and the target category class. Attributes such as Total Rainfall, Week, Month, and Crude Protein, have less of a correlation as the values are closer to 0 than 1.

Information Gain Attribute Evaluation with a ranker filter was also used for feature selection. In this method, attributes such as Date, Offtake, Dry Matter, Available, and Pre-Grazing Cover, provide more information to the prediction of the target class. Attributes which provide less information for predicting included Acid Detergent Fiber, Dry Matter, and Post Grazing Cover, while attributes such as Month, WSC, Utilisation, Crude Protein, Air Temperature, Total Rainfall, and Post-Grazing Cover, had no information gain with a value of 0.

Another method of feature selection used on the data was the Wrapper Subset Evaluation, using a Decision Tree with the Best First Ranker method. This method uses a decision tree to evaluate numerous subsets to determine the best subset. In this method, the merit of the best subset was 0.748, and found that the optimal number of folds for this dataset is 5 folds. Attributes identified as having the most significance included Week, ADF, ME, Utilisation and Total Rainfall.

The feature selection methods described above have selected different attributes as being the most informative. Each of the classification methods were run again using the information provided by the feature selection methods. However, the results proved to be poorer when features were removed. A Python script was developed to perform the same feature selection methods, and difficulties were found due to the text fields (County, Event, Date) and negative numbers (Utilisation, Offtake, Available) in the dataset, which are problematic in the feature selection methods chosen. The negative numbers were normalised and the text fields were categorised, and the classification was run again. As it had been before, each of the methods chose different features as important and there was no strong similarity between features. As well as this, features which would be designated important in real life (*e.g.*, rainfall) were not classed very highly and vice versa. Therefore, some further investigation on the feature selection methods to better suit the available dataset needs to be done, along with the more elaborate data pre-processing method to be used to classify and clean the data in order for the above feature selection methods become feasible.

## 5 Conclusions and Future Work

The present work focused on analysing and evaluating four data-driven classifiers for grass growth prediction using some real grass data collected related to soil content, weather, grass quality components *etc.* From the above study, it was found that tree classifiers were better methods of classification, namely the DT and RF methods. DT performed better in datasets which contained imbalance, such as in the seasonal divisions of spring, summer, and autumn. It consistently performed the best in Kappa statistics, precision, recall and F-measure across all seasonal data. RF performed consistently in RMSE and ROC area in both imbalanced and balanced datasets, with its best performance values as low as 0.27 (RMSE), and as high as 0.96 (ROC) on the

spring dataset with synthetic data. This was the highest ROC value across all the classifiers, while the lowest value was 0.69, produced by NB on the autumn dataset. Once synthetic data were applied, and the imbalance was eradicated, RF became the overall best classifier in each of the experiments that was carried out. It consistently performed the best in Kappa, RMSE, precision, recall, F-measure, and ROC area. NB could be considered the least successful classification method, as its accuracy and evaluation metrics were well below that of the other three methods. The best performance from this classifier was on the spring dataset with SMOTE applied, in which its accuracy was 78%, while the other classifiers were 84% and 85% accurate in the same dataset. Other measures including MAE (0.16) were good in the NB classifier, however, this was the highest error rate in the dataset. Other interesting results were produced by the DT classifier, as it reduced in performance, over all of the evaluation metrics, from the whole year to the dataset when instances containing missing data were removed. This is due to the size of the dataset, as a small dataset of 107 instances does not contain enough information for the DT to make accurate decisions. The above study has demonstrated the good potential of using data analytics for grass growth prediction, although the overall performance of those four classifiers are not exceptional considering, for example, the accuracy rate, which is partially due to the quality of data (missing data, imbalance and uncertainty inside), and partially due to limiting to only four classifiers. More elaborate data preprocessing and cleaning methods can be used, and other types of classifiers can be also explored further in future work. One limitation of this study is the inability to explain how the classifiers came to the conclusion of their prediction. At present the classifiers are assigned greater weights when there is higher information gain, and lower weights when there is little information gain. Future work will consider the use of expert knowledge to assign weights to attributes, which will allow the conclusion to be better explained to users, and to give definitive reasoning for the prediction.

This study underpins research for aiding farmers and policy makers in their decisions regarding sustainable land management. The agriculture and farming industry of NI requires tools and strategies to encourage sustainable land management, especially due to its greater contribution to gaseous emissions in NI. The study has highlighted the need for a system that can handle missing data and uncertainty. The data-driven approach is expected to be combined with expert knowledge from the industry and models must be integrated to enhance the overall performance and create a multilayer decision support system, to support farmers and policymakers when making land sustainability decisions.

## References

- [1] Parliament of the United Kingdom, "Climate Change Act 2008," 2008.
- [2] Committee on Climate Change, "Reducing emissions in Northern Ireland," 2019.
- [3] J. McDonnell, C. Brophy, E. Ruelle, L. Shalloo, K. Lambkin and D. Hennessy, "Weather Forecasts to Enhance an Irish Grass Growth Model," *European Journal of Agronomy*, vol. 105, pp. 168-175, 2019.

- [4] O. McHugh, F. Browne, J. Liu and P. Jordan, "A Decision Analytic Framework and Exploratory Statistical Case Study Analysis of Grass Growth in Northern Ireland," *Journal of Advances in Information Technology*, vol. 11, no. 1, pp. 15-20, 2020.
- [5] E. Ruelle, D. Hennessy and L. Delaby, "Development of the Moorepark St Gilles grass growth model (MoSt GG model): A predictive model for grass growth for pasture based systems," *European Journal of Agronomy*, pp. 80-91, 2018.
- [6] U. Ayub and S. A. Moqurrab, "Predicting Crop Diseases Using Data Mining Approaches: Classification," in *1st International Conference on Power, Energy and Smart Grid (ICPESG)*, 2018.
- [7] D. Ramesh and B. V. Vardhan, "Analysis of Crop Yield Prediction Using Data Mining Techniques," *IJRET: International Journal of Research in Engineering and Technology*, vol. 04, no. 01, pp. 470-473, 2015.
- [8] I. Ali, F. Cawkwell, E. Dwyer and S. Green, "Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 07, pp. 3254-3264, 2017.
- [9] D. C. Corrales, J. C. Corrales and A. Figueroa-Casas, "Towards Detecting Crop Diseases and Pest by Supervised Learning," *Ingeniería y Universidad*, vol. 19, no. 01, pp. 207-228, 2015.
- [10] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.