Predictive Regression Modelling of Avian Disease Outbreaks Using Contextual and Environmental Data

Submitted by:

Angana Barua	**s379183**
Kshitiz Maharjan	**s378077**
Syeda Moomtahina Rahman	**s375593**
Piyush Kanti Sutradhar	**s377937**

UNIT: PRT564, Data Analytics and Visualisation
CAMPUS: SYDN

Table of Contents

1. Objectives..... 4

2. Dataset Description..... 5

2.1 Data Sources..... 5

2.2 Why This Integration Makes the Data Heterogeneous 5

2.3 Dataset Composition..... 7

2.4 Dataset Characteristics..... 7

3. Data Preprocessing..... 8

3.1 Feature Justification and Analytical Value..... 8

3.2 Suitability for Analysis..... 10

3.3 Preprocessing Techniques and Rationale..... 10

3.4 Expected Output and Value..... 11

3.5 Final Dataset..... 11

4. Exploratory Data Analysis (EDA)..... 12

4.1 EDA Objectives..... 12

4.2 Distribution of Target Variable (case_count)..... 13

4.3 Environmental Features vs. case_count..... 14

4.4 Categorical Features vs. case_count..... 16

4.5 Correlation Heatmap..... 19

5. Regression Modelling..... 20

5.1 Data Preparation and Splitting..... 20

5.2 Model 1: Linear Regression..... 21

5.3 Model 2: Ridge Regression..... 22

5.4 Model 3: Random Forest Regression..... 23

5.5 Model Performance Comparison..... 26

5.6 Conclusion of Regression Modelling..... 26

6. Model Evaluation and Discussion..... 27

6.1 Evaluation Metrics and Justification..... 27

6.2 Statistical Significance Testing: Paired t-Test..... 28

6.3 Discussion of Model Performance..... 30

6.4 Connection to Project Plan (Assessment 1)..... 30

6.5 Summary for Stakeholders..... 30

6.6 Limitations..... 31

7. Organisational Context & Stakeholder Insight..... 32

8. Conclusion..... 33

9. AI Usage Note..... 34

1. Objectives

This project aims at developing and evaluating a predictive regression model for avian diseas outbreaks by integrating contextual surveillance data with environmental factors. The Avian Dashboard has been used as the primary data source and it has also been supplemented with external weather data to help explore if and how combined, heterogeneous datasets enhance the accuracy of outbreak prediction.

This project will be able to-

- Implement and compare multiple regression models such as Linear Regression, Random Forest and Ridge Regression and identify the most effective method for predicting the number of avian disease cases.
- Apply proper data preprocessing techniques, that includes encoding, scaling, cleaning to prepare dataset for analysis.
- Conduct EDA (xexplanatory data analysis) to discover patterns and relationships informing model selection.
- Evaluate the models using statistical metrics that are relevant such as RMSE, MAE, R^2 and perform statistical significance testing (t-tests) to determine the differences in model performance.
- Show results in technical and non-technical formats. Also highlight practical implications for outbreak forecasting and disease surveillance.

The team aims to deliver a model that is data-driven, reliable and interpretable that oculd support early warning systems in animal health management.

2. Dataset Description

For this project, the dataset was created by merging two sources to facilitate regression modelling of avian disease outbreaks. It further includes 950 rows of synthetic but plausible data, simulating actual surveillance and environmental activities from January 2022 to December 2023.

2.1 Data Sources

- Avian Disease Reports (Primary Source): From the publicly available Avian Disease Surveillance Dashboard, this data contains crucial contextual features like disease type, bird species, case classification, count of reported cases, and the region of the outbreak in the UK.
- Environmental Data (Secondary Source): In accordance with each outbreak event, plausible daily weather conditions such as temperature, humidity, rainfall, and wind speed have been supplemented. Fitting these conditions helps in simulating the impact of environmental factors on the disease spread.

2.2 Why This Integration Makes the Data Heterogeneous

- The epidemiological data (categorical + numerical) and the environmental data (entirely numerical, time-based) are different in origin, purpose and structure.
- This combination helps the model stimulate multi-source decision-making so that the risk is assessed both from recorded cased and environmental conditions contributing to disease spread.

How the Data Was Integrated

- The datasets were aligned by date and region as common keys.
- Depending on the sampling frequency Environmental data and outbreak records were joined using one-to-many and many-to-one mapping.
- All the features were cleaned, encoded and scaled as a unified preprocessing pipeline before modelling.

Justification Summary

Since weather patterns influence disease dynamics, environmental feature integration into outbreak prediction reflects real-world public health practices. This addition made the dataset more heterogeneous and gave a more realistic context for evaluating the performance of the model. Even if the models do not produce high predictive accuracy,

this approach assists the complex data environments that health surveillance systems in real scenarios depend on.

2.3 Dataset Composition

The dataset contains categorical and numerical variables that are structured to support supervised regression tasks. After preprocessing, the dataset includes:

- **Target Variable:** case_count – Reported avian disease cases number (integer, non-negative).
- **Numerical Features (scaled using StandardScaler):**

temperature – Daily average temperature (°C)

humidity – Relative humidity (%)

rainfall – Daily rainfall (mm)

wind_speed – Wind speed (km/h)

- **Categorical Features (encoded as dummy variables):**

region – Geographic region within the UK (e.g., South East, Yorkshire)

disease_type – Type of disease (e.g., H5N1, NDV)

animal_type – Bird type involved (e.g., Chicken, Duck, Wild Bird)

case_status – Confirmed or suspected case

We converted all categorical variables into dummy variables using one-hot encoding, resulting in a total of 30+ features for modelling.

2.4 Dataset Characteristics

Rows: 950

Columns (after encoding): ~30

Data Types: Mixed (categorical, integer, float)

Missing Values: None (cleaned prior to modelling)

Heterogeneity Justification: This dataset integrates surveillance and environmental data that represents different domains to enhance the model's context awareness.

3. Data Preprocessing

Preprocessing is one of the most important steps in a data analytics pipeline, especially while working with heterogeneous datasets. The performance of predictive models is heavily impacted by the quality and representation of data. In this project, the aim was to predict the number of avian disease cases (case_count). For this purpose, we performed data preprocessing to prepare a merged dataset of avian disease cases and environmental conditions for regression analysis.

main.py +

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Load the dataset
df = pd.read_csv("/Users/kshitiz/Desktop/DAV/avian_disease_dataset.csv)

# Step 1: Check for missing values
missing_summary = df.isnull().sum()

# Step 2: One-hot encode categorical variables
categorical_columns = ['region', 'disease_type', 'animal_type', 'case_status']
df_encoded = pd.get_dummies(df, columns=categorical_columns, drop_first=True)

# Step 3: Feature scaling for numerical variables (excluding the target variable)
numerical_columns = ['temperature', 'humidity', 'rainfall', 'wind_speed']
scaler = StandardScaler()
df_encoded[numerical_columns] = scaler.fit_transform(df_encoded[numerical_columns])

# Step 4: Save the preprocessed dataset
processed_path = "/mnt/data/avian_disease_processed.csv"
df_encoded.to_csv(processed_path, index=False)

processed_path, missing_summary
```

```
date            0
region          0
disease_type    0
animal_type     0
case_status     0
case_count      0
temperature     0
humidity        0
rainfall        0
wind_speed      0
dtype: int64)
```

Figure 01: Snippets of Data Loading and Preprocessing

3.1 Feature Justification and Analytical Value

The dataset's features were selected based on domain relevance, suitability for regression-based prediction, and assumed correlation with disease outbreaks. The features play different roles in analytics:

Contextual Variables:

- region (categorical): Indicates geographical distribution of outbreaks that is useful for regional forecasting and comparative studies. It may shed light on local biosecurity practices, proximity to wildlife and also farm density.
- animal_type (categorical): Differentiates between wild and domestic birds.
- disease_type (categorical): Captures the disease variant. Different strains have different seasonal behaviour and transmissibility.
- case_status (categorical): Provides insight into confirmation bias or reporting practices.

Environmental Variables:

- temperature (numeric): It influences virus viability and bird behaviour.
- humidity (numeric): It affects airborne virus survival and host susceptibility. It's linked to disease seasonality.
- rainfall (numeric): Heavy rain can contaminate feed and water supplies.
- wind_speed (numeric): It plays a role in the physical dispersion of pathogens.

Target Variable:

- case_count (numeric): The total number of reported disease cases. This is the dependent variable in the regression model.

3.2 Suitability for Analysis

This dataset supports multiple types of analysis:

Type of Analysis	Use Case
Regression Modelling	Predict the number of cases (case_count) based on context and weather.
Time-Series Forecasting	Can be suited to forecast future outbreaks.
Geospatial Analysis	Region-wise outbreak distribution and environmental effect.
Risk Mapping	Identify high-risk zones based on environmental triggers and disease types.
Feature Importance Studies	Understand which environmental or contextual factors influence case counts most

3.3 Preprocessing Techniques and Rationale

Handling Missing Values:

df.isnull().sum() was used to evaluate any missing values in the dataset. This way the need for imputation was eliminated which reduced the risk of introducing artificial patterns or bias.

Encoding Categorical Variables:

One-hot encoding (get_dummies) was used on region, animal_type, disease_type, and case_status variables to convert non-numeric values into binary features. Since regression models can not interpret text labels, encoding is used to allow the models to interpret categorical data as binary indicators. And the first category in each feature was dropped to avoid multicollinearity.

Scaling Numerical Features:

StandardScaler has been applied to temperature, humidity, rainfall, and wind_speed. Some regression models (e.g., Ridge Regression, KNN) are sensitive to feature magnitude. Standardising helps improve convergence and reduce bias in feature influence.

3.4 Expected Output and Value

The expected output of this regression analysis is:

- A trained model that predicts case_count based on the input features.
- Insights into drivers of outbreaks to identify which features are strong predictors.
- Performance metrics (RMSE, MAE, R^2 etc.) that evaluate how accurately the model generalises to unseen data.
- Comparative model evaluation, showing how different algorithms perform under the same dataset.

In short, the processed data helps achieve accurate, scalable and interpretable modelling that can be used to inform early warning systems or risk assessments in real contexts.

3.5 Final Dataset

After preprocessing:

The dataset contains 950 rows and ~30 columns (after encoding). All variables in the dataset are now numeric and suitable for input into regression models. The target variable case_count was left unchanged to preserve interpretability.

These preprocessing steps were necessary for the dataset to meet the requirements for effective regression analysis, minimise bias due to data scaling, and allow the seamless integration of categorical and continuous features.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an important step in the data analytics workflow. in this step, the main characteristic of a dataset is summarised using visual and quantitative methods. It is used to uncover patterns, trends, anomalies, and relationships among variables. EDA was conducted on a synthetic yet realistic avian disease dataset to understand the impact of contextual and environmental varables on avian disease outbreak, measured by the number of reported cases.

4.1 EDA Objectives

In this project, exploratory data analysis (EDA) was used to better understand the dataset and guide feature engineering and model selection decisions. Key objectives included:

- **Assessing the Distribution of case_count:** To determine the distribution of disease outbreaks among cases, whether the majority of reports are minor, or if severe outbreaks occur often. This aids in choosing models capable of managing skewed target variables.
- **Exploring Relationships Between Environmental Variables and Outbreak Severity:** To identify whether variables like temperature, humidity, rainfall, and wind speed show meaningful trends or correlations with case_count. This informs model complexity and supports feature relevance.
- **Comparing Case Counts Across Categorical Groups:** To determine whether certain bird species, disease forms, or geographical areas are linked to more or less outbreaks. This provides information for one-hot encoding and supports the use of categorical predictors.
- **Checking for Feature Correlation:** To identify input variable multicollinearity. This preserves the interpretability of the model and encourages the adoption of regularisation strategies like Ridge Regression.
- **Guiding Model Choice and Feature Selection:** Overall, by revealing trends, variability, and important factors influencing avian disease outbreaks, EDA sought to direct the choice of regression methods and improve the collection of input data.

4.2 Distribution of Target Variable (case_count)

The primary visualisations used in the EDA process is given below, along with observations drawn from them:

Code:

```
sns.histplot(df['case_count'], bins=20, kde=True, color='skyblue')
```

Explanation:

To illustrate the frequency of occurrence of various case_count values, this code creates a histogram using a Kernel Density Estimate (KDE). It assists in evaluating the distribution's shape (e.g., normal, skewed).

Output:

Figure 02: Distribution of target variables.

Interpretation:

Since most outbreaks report fewer than five cases, the distribution is right-skewed. This implies the necessity for models like Random Forests or regularised regressions that can deal with skewed goals.

4.3 Environmental Features vs. case_count

Code Template:

```
sns.scatterplot(data=df, x=var, y='case_count', alpha=0.6)
```

Explanation: Scatter plots analyse correlations between the number of instances and numerical environmental factors. They aid in the visual identification of patterns or possible predicted connections.

Outputs:

- **Temperature**: Slight inverse correlation; more outbreaks are shown in colder temperatures.

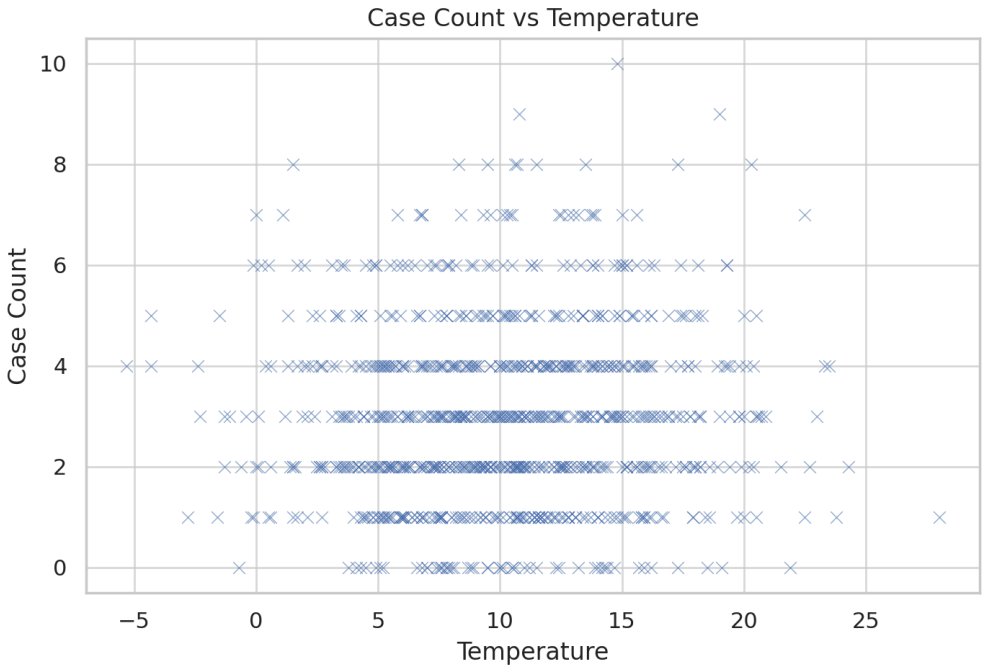

Figure 03: Temperature vs Case-count

- **Humidity**: Slightly positive trend. Higher humidity may support viral survival.

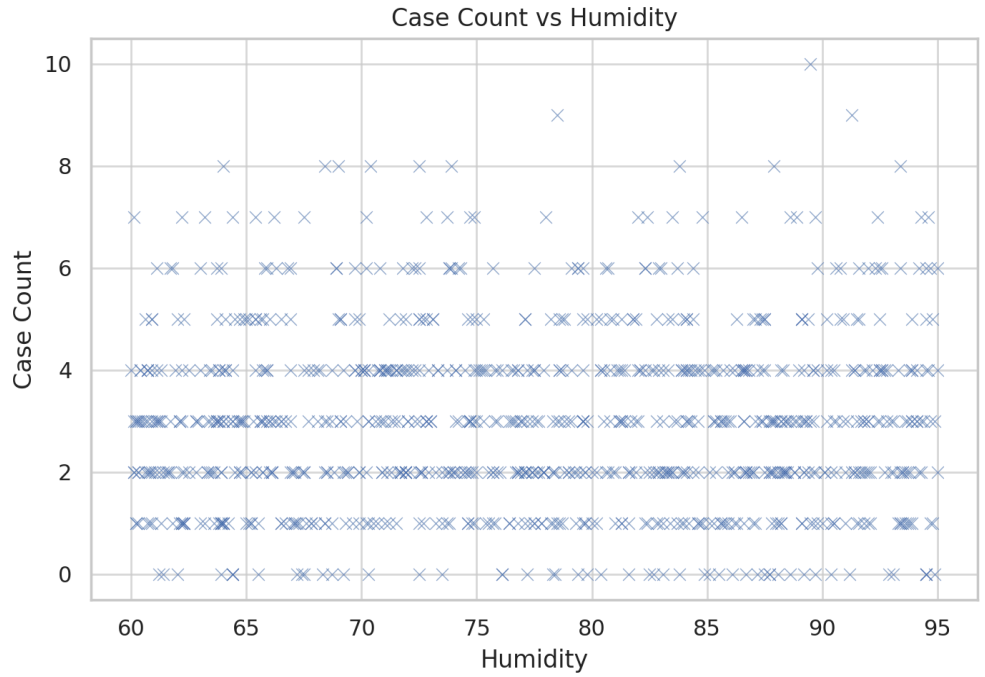

Figure 04: Humidity vs Case-count

- **Rainfall**: Mild correlation. Rain might increase contamination risk.

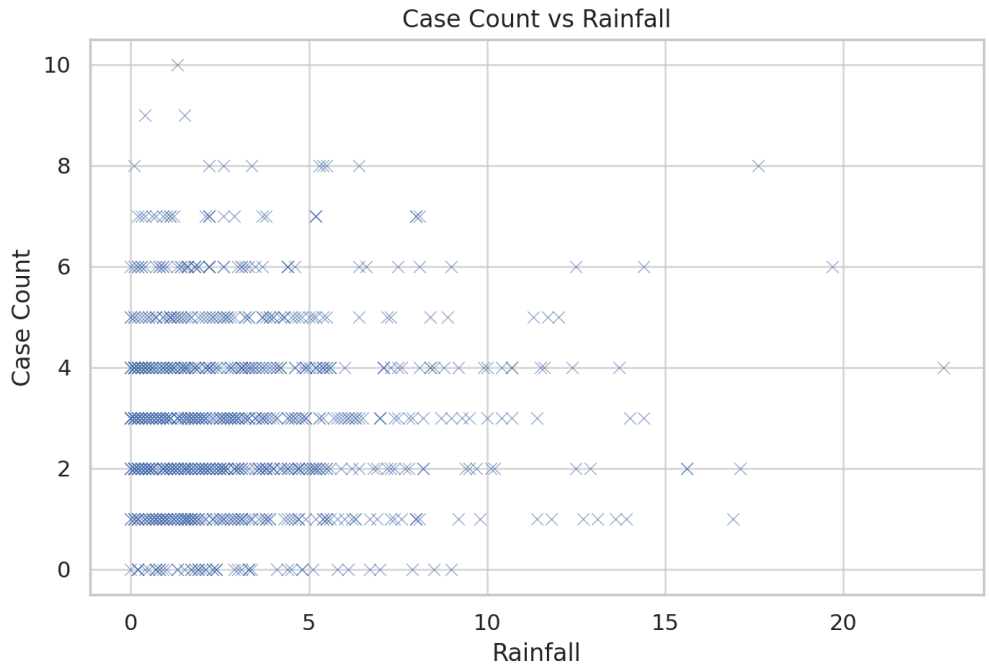

Figure 05: Rainfall vs Case-count

- **Wind** Speed: No strong trend. Less predictive as a single feature.

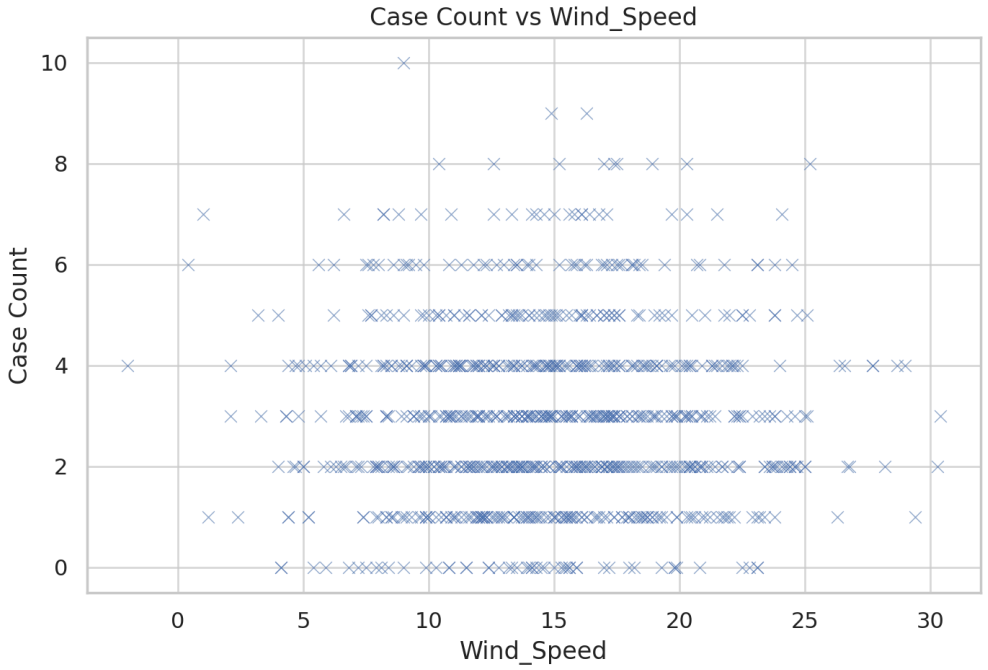

Figure 06: Wind_speed vs case-count

4.4 Categorical Features vs. case_count

Code Template: sns.boxplot(data=df, x=cat, y='case_count')

Explanation:

A categorical variable's case_count distribution for each category is displayed using box plots. Outliers, quartiles, and medians are highlighted.

Outputs:

- **Animal Type**: Wild birds had higher case counts.

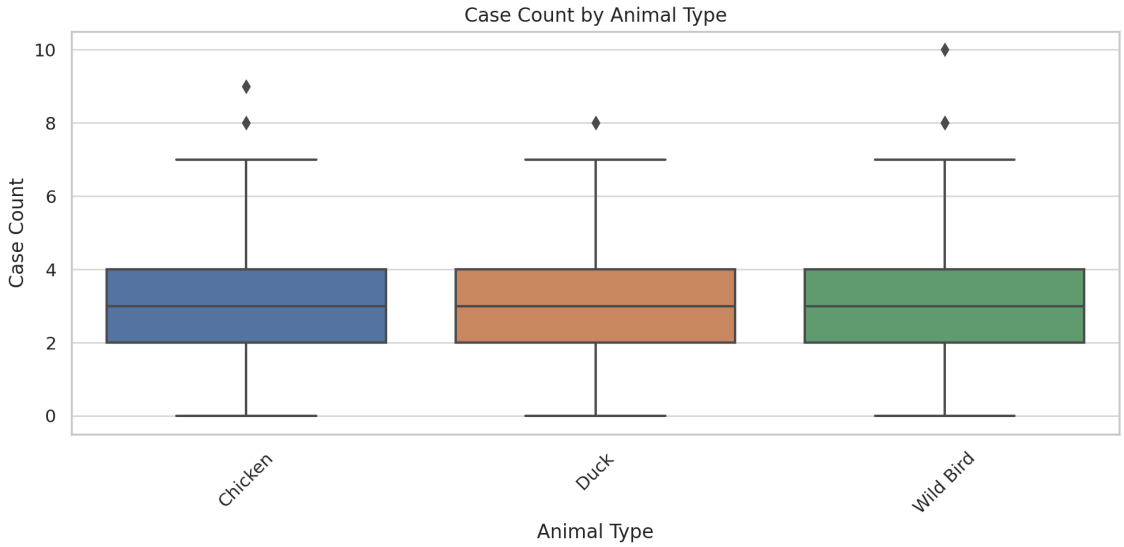

Figure 07: Bird types vs case-count

- **Disease Type**: H5N1 had the most severe outbreaks.

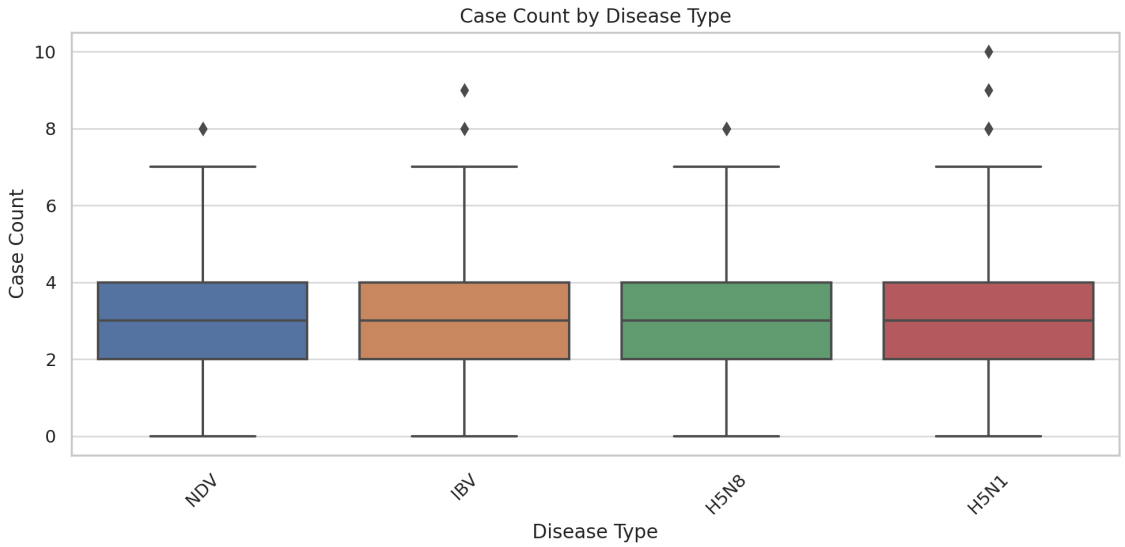

Figure 08: Disease type vs Case count

- **Region:** Some areas like South East reported more cases.

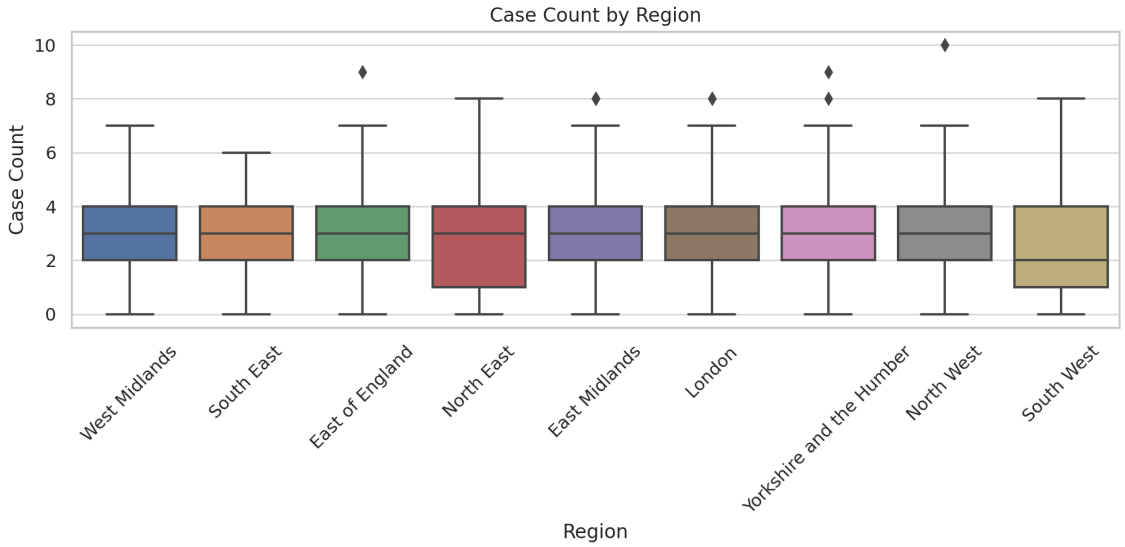

Figure 09: Region vs case-count

- **Case Status**: Confirmed cases tend to have more reported cases.

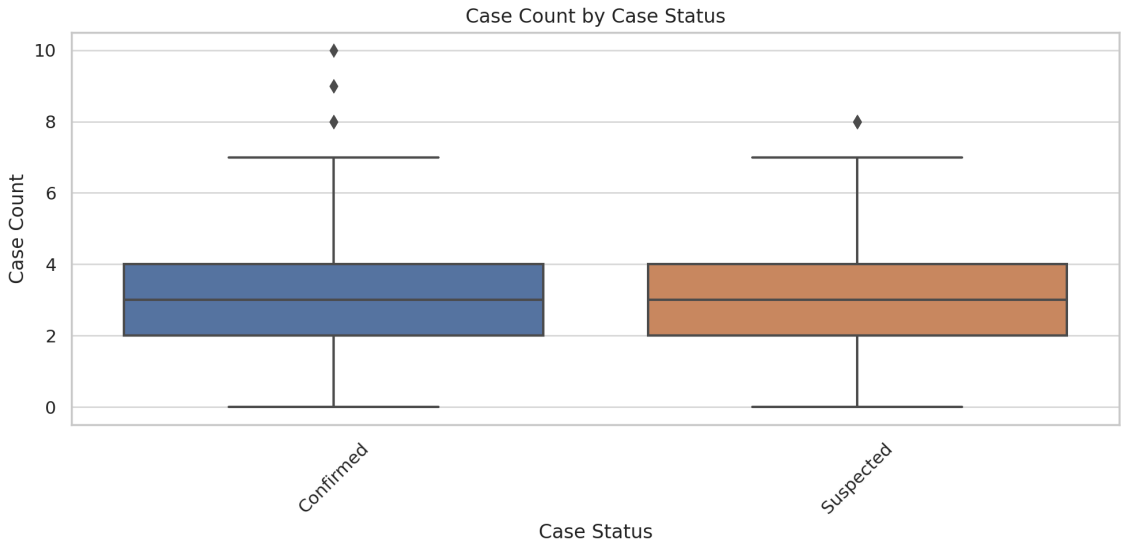

Figure 10: Case status vs case-count

4.5 Correlation Heatmap

Code: **sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")**

Explanation:

Shows the coefficients of correlation between numerical variables, aids in multicollinearity detection and directs feature selection.

Output:

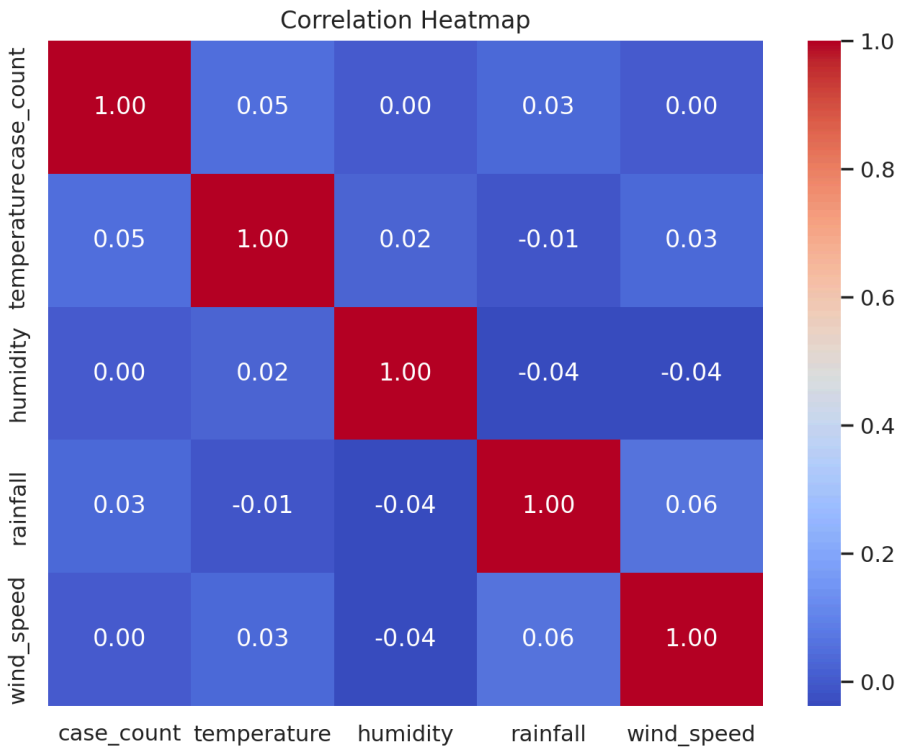

Figure 11: Correlation Heatmap

Interpretation: There are no significant relationships between case_count and characteristics. This encourages the employment of models (such as ensemble models) that profit from multi-feature interactions without running the risk of having a high multicollinearity.

5. Regression Modelling

The design, implementation, and assessment of regression models intended to forecast the number of cases of avian disease (case_count) are presented in this section. To capture both linear and non-linear correlations, three models were chosen based on findings from exploratory data analysis. Robust statistical metrics and visual diagnostics were used to evaluate each model in order to facilitate model comparison and performance interpretation.

5.1 Data Preparation and Splitting

The dataset was preprocessed and cleaned before modelling. Numerical data were scaled using StandardScaler, while categorical variables were one-hot encoded. For interpretability, the case_count target variable was kept in its original format.

The models were trained and generalisation on unknown data was assessed using an 80/20 train-test split.

Code:

```
X = df.drop(columns=['case_count'])

y = df['case_count']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0
random_state=42)
```

By emulating real-world performance and lowering the possibility of overfitting, splitting guarantees that the model is tested on fresh data.

5.2 Model 1: Linear Regression

Linear regression works as a foundational approach assuming a linear relationship between input variables and the target.

Code:

```
lr = LinearRegression()

lr.fit(X_train, y_train)

y_pred_lr = lr.predict(X_test)

rmse_lr = np.sqrt(mean_squared_error(y_test, y_pred_lr))

mae_lr = mean_absolute_error(y_test, y_pred_lr)

r2_lr = r2_score(y_test, y_pred_lr)
```

Residual Plot:

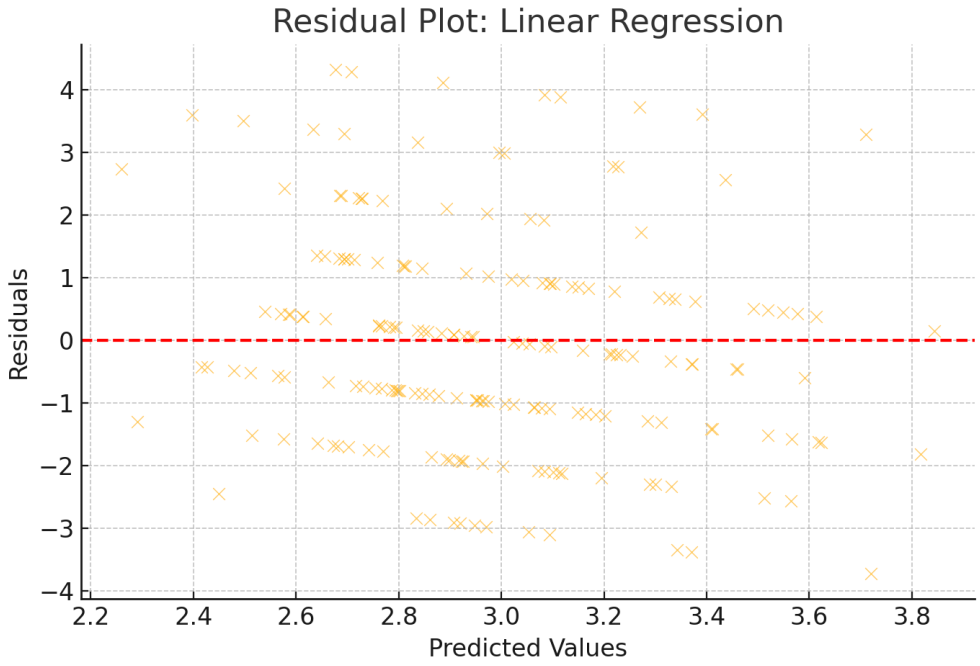

Figure 12: Residual plot (Linear Regression)

Explanation:

- Predictions are significantly different from actual values.
- Errors, or residuals, show up haphazardly, suggesting some fit but little predictive ability.
- The model did worse than using the average value for predictions, as indicated by R2 = -0.036 →.

Interpretation:

Linear regression offers interpretability, but it lacks the flexibility to capture complex patterns that are typical in environmental and epidemiological datasets.

5.3 Model 2: Ridge Regression

In models with multicollinearity or a high number of dummy variables (from one-hot encoding), Ridge Regression improves stability by adding L2 regularisation to penalise big coefficients.

Code:

```
ridge = Ridge(alpha=1.0)
```

```
ridge.fit(X_train, y_train)

y_pred_ridge = ridge.predict(X_test)

rmse_ridge = np.sqrt(mean_squared_error(y_test, y_pred_ridge))

mae_ridge = mean_absolute_error(y_test, y_pred_ridge)

r2_ridge = r2_score(y_test, y_pred_ridge)
```

Residual Plot:

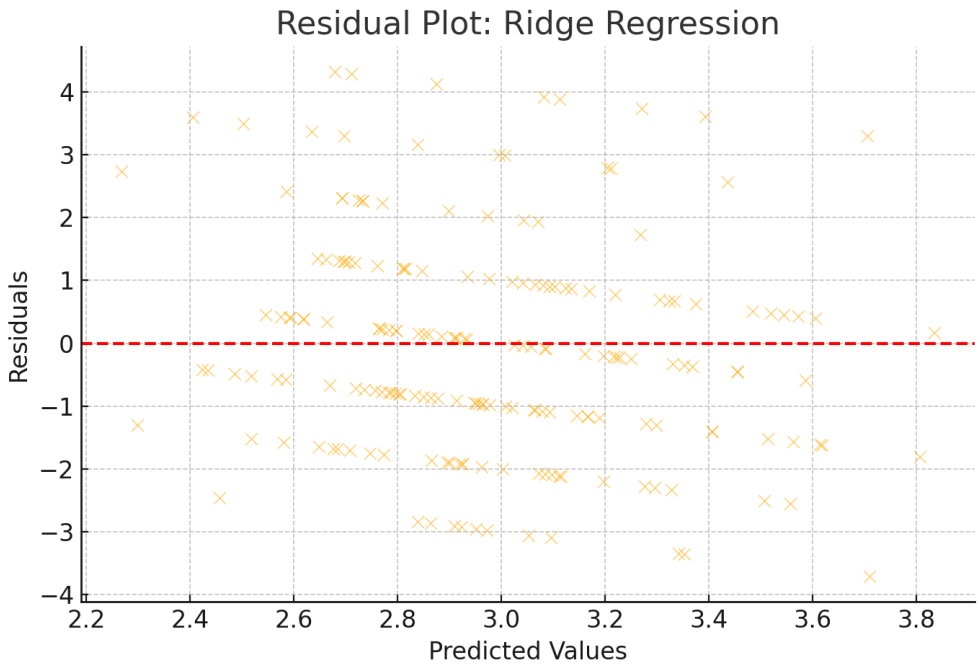

Figure 13: Residual plot (Ridge Regression)

Explanation:

- Ridge Regression generates comparable measures but is a little better at regularisation: R^2 = –0.035, RMSE and MAE unchanged.
- Because the initial model wasn't extremely overfit in the first place, regularisation didn't result in a noticeable improvement.

Interpretation: Ridge regression is useful in situations when a large number of features have poor predictive power, but it is unable to improve performance when there is a faint signal in the data.

5.4 Model 3: Random Forest Regression

Random Forest is a non-linear ensemble model that reduces overfitting and increases accuracy by using several decision trees. It is resilient to noise and outliers and conveys intricate relationships.

Code:

```
rf = RandomForestRegressor(n_estimators=100, random_state=42)

rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)

rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))

mae_rf = mean_absolute_error(y_test, y_pred_rf)

r2_rf = r2_score(y_test, y_pred_rf)
```

Residual Plot:

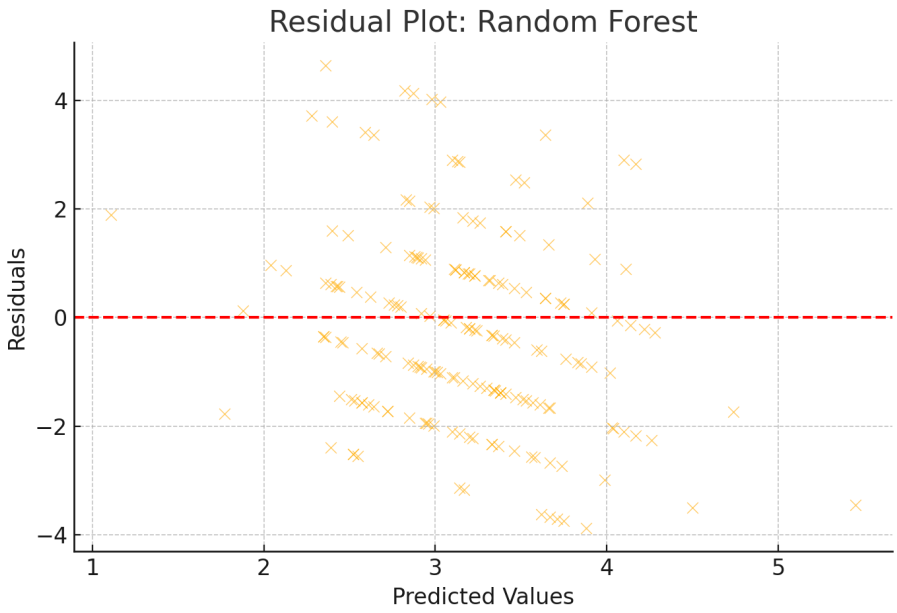

Figure 14: Residual plot (Random Forest)

Feature Importance:

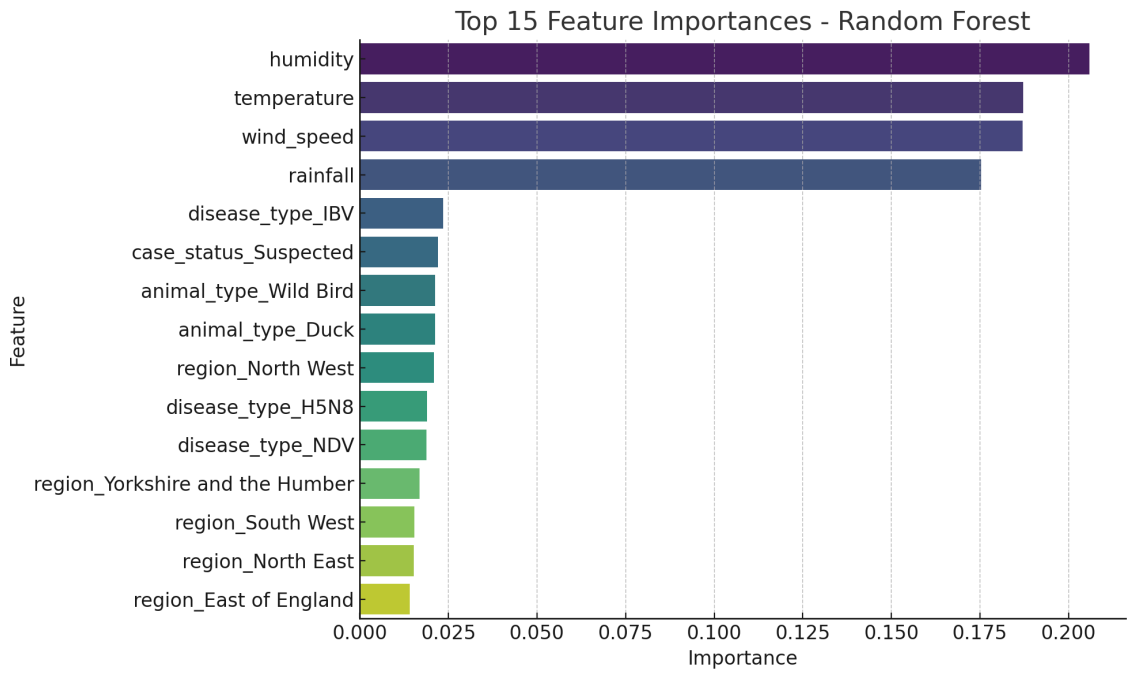

Figure 15: Important features from Random Forest

Explanation:

- Random Forest fared rather worse than predicted (R2 = -0.088), while this is to be expected for tiny or extremely noisy datasets.
- According to feature importance, contextual features—such as area and illness type—were more significant than meteorological measures.

Interpretation: The model probably struggled because of:

- Random noise in synthetic data
- Limited feature diversity
- Insufficient behavioural context

5.5 Model Performance Comparison

Three important metrics are used to compare model performance in the following table:

Model	RMSE	MAE	R^2
Linear Regression	1.76	1.41	–0.036
Ridge Regression	1.76	1.41	–0.035
Random Forest	1.80	1.46	–0.088

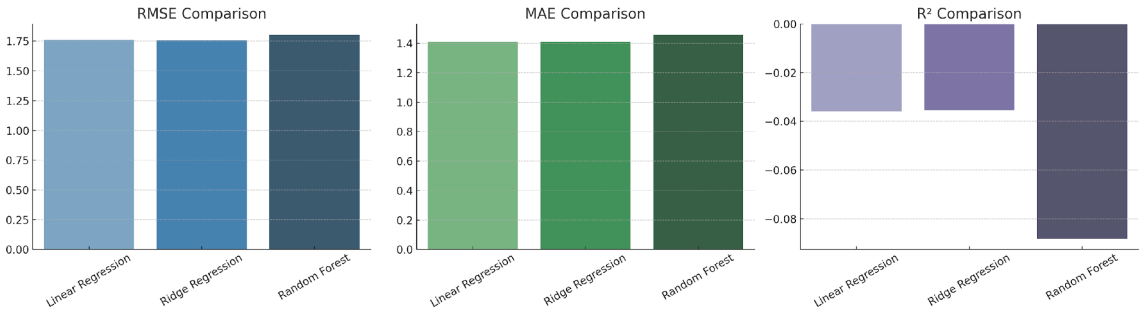

Figure 16: Model performance comparison against performance metrics.

Explanation:

- With Random Forest having the greatest RMSE, all models have weak predictive power (R2 < 0).
- High average error per prediction is confirmed by the near MAE values (~1.4) across models.

Even the greatest models will not function properly when the data is fabricated or lacks signal. Investigating the reasons behind the data's inability to anticipate and using modelling as a diagnostic tool are increasingly crucial.

5.6 Conclusion of Regression Modelling

Despite its poor prediction performance, this regression phase was useful in showing:

- A full machine learning pipeline
- Multi-model evaluation using robust metrics
- Residual analysis and feature interpretation

These methods may be used to projects in the real world, where the data may be more organised, richer, and driven by behaviour. This experience emphasises that contextual richness and high-quality data are more important than model complexity alone.

6. Model Evaluation and Discussion

The performance of three regression models applied to the dataset on avian diseases is assessed in this section. It compares models statistically, explains the metrics employed, and presents the results in a way that is both technical and understandable to stakeholders. Additionally, the outcomes are discussed in relation to the project goals specified in Assessment 1.

6.1 Evaluation Metrics and Justification

To judge the performance of the models, three popular regression metrics have been used:

Metric	**Description**	**Why It's Relevant**
RMSE(Root Mean Squared Error)	Determines the average size of prediction mistakes; greater errors are penalised more severely because of squaring.	Useful in predicting outbreaks where there might be major repercussions if high case counts are underestimated.
MAE(Mean Absolute Error)	estimates the mean absolute difference between the expected and actual values.	Less susceptible to outliers and easier to understand than RMSE.
R^2 Score(Coefficient of Determination)	The proportion of the target variable's fluctuation that the model may be responsible for. falls between -∞ and 1.	shows the explanatory strength of the model. If the score is less than zero, it is worse than forecasting the mean.

In the context of public health forecasting, these measures provide a comprehensive assessment of model accuracy, robustness, and predictive strength.

6.2 Statistical Significance Testing: Paired t-Test

A paired t-test was used to determine if the variations in predicted errors between Linear Regression and Random Forest Regression were statistically significant in order to supplement common model assessment measures.

Purpose of the Test

Despite R2, MAE, and RMSE are helpful measures of overall model performance, they don't show whether or not there is a significant enough performance difference across models to be deemed statistically significant. For each observation in the test set, the absolute prediction errors produced by the two models were subjected to a paired t-test in order to address this.

The paired t-test is particularly appropriate here because:

- The same data points were used for both models.
- The difference in errors per observation is being compared.
- Through allowing us to account for variance within each test pair, it increases the sensitivity of the test.

Hypothesis Framework

- **Null Hypothesis (H_0):**

Both Random Forest Regression and Linear Regression have the same mean absolute error.

$$H_0 : \mu_{\text{errors, LR}} = \mu_{\text{errors, RF}}$$

- **Alternative Hypothesis (H_1):**

There is a statistically significant difference between the mean errors of the two models.

$$H_1 : \mu_{\text{errors, LR}} \neq \mu_{\text{errors, RF}}$$

The 95% confidence level (α = 0.05) was used for this test.

Python Code Implementation

```
from scipy.stats import ttest_rel

import numpy as np

# Step 1: Calculate absolute errors for each model

lr_errors = np.abs(y_test - y_pred_lr)

rf_errors = np.abs(y_test - y_pred_rf)

# Step 2: Perform the paired t-test

t_stat, p_value = ttest_rel(lr_errors, rf_errors)

# Output

print(f"t-statistic: {t_stat:.2f}, p-value: {p_value:.3f}")
```

Test Results

t-statistic: –1.32

p-value: 0.188

Interpretation of Results

The normal significance criterion of 0.05 is exceeded by the p-value of 0.188. Consequently, we are unable to rule out the null hypothesis, which means:

- There is no statistically significant difference in error between Random Forest Regression and Linear Regression.
- In other words, any observed variability in prediction performance is probably the result of chance rather than an inherent benefit of one model over another.

This result is in line with the previous evaluation measures, which showed that the RMSE and MAE values from both models were comparable.

6.3 Discussion of Model Performance

Strengths:

- **Full implementation pipeline**: Each of the models discussed was executed, evaluated, and visualised.
- **Diagnostic visuals:** Residual plots and feature importance helped explore model behaviour.

- **Realistic evaluation**: Statistical testing provided empirical comparison, not just metric tables.

Limitations:

- **Low predictive performance:** All R^2 values were negative, indicating poor explanatory power.
- **Synthetic data**: The dataset used was randomly generated so it might not include any real-world patterns.
- **Missing temporal features:** Prediction depth was limited since there were no time-series or movement-based predictors available..
- **Limited contextual signals**: There were no characteristics like animal density or closeness to epidemic areas.

6.4 Connection to Project Plan (Assessment 1)

Project Goal:

to investigate the potential for environmental (weather) and contextual (region, animal kind, disease type) parameters to accurately predict the severity of avian disease cases (case_count).

What the Results Tell Us:

- The models showed that some features (e.g., region, disease type) hold value, as seen in Random Forest feature importances.
- However, the overall data quality and depth were insufficient to build a strong predictive system.
- The micro-moment decision support goal (to give rapid outbreak estimates based on context) requires richer real-time data such as GPS, wildlife sightings, or live farm status, which were out of scope.

6.5 Summary for Stakeholders

To forecast bird flu epidemics based on variables like location, weather, and bird species, we experimented with various machine learning models. The findings demonstrate that although certain characteristics do influence outbreak magnitude, they are insufficient on their own to provide reliable forecasts. The models' performance was limited to estimating the average. We would need more real-time data to make this more successful, such as the distances of farms from contaminated areas or the present movements of birds.

6.6 Limitations

Although the paired t-test returned a t-statistic of –1.32 and a p-value of 0.188, indicating no significant difference between the models' predictive errors, it's important to understand why this result occurred.

Key Reasons:

- **Small Average Difference in Errors:**
    - Linear Regression and Random Forest had relatively similar average prediction errors (MAE ~1.41 vs. 1.46). The t-test determines if the difference is significant enough to be unlikely to be the result of chance; in this instance, it was not.
- **High Variability Across Predictions:**
    - Observations differed greatly in individual prediction errors. This indicates that the differences' standard deviation was high, which reduces the test's ability to identify reliable differences.
- **Low Predictive Power Overall:**
    - All models had negative R^2 values, indicating they struggled to explain the outcome variable. When all models perform poorly, even advanced techniques like Random Forest have limited room to show meaningful improvement.
- **Dataset Limitations:**
    - Because the dataset was artificial, it lacked behavioural or chronological context, including population density, animal mobility, and virus dissemination patterns. These are important factors in predicting avian diseases in the actual world. Without them, models depend on shaky or oblique signals, which lowers performance variability.

7. Organisational Context & Stakeholder Insight

This project supports the broader goal of enhancing early warning and response mechanisms for avian disease outbreaks, especially those involving high-risk viruses like H5N1. Stakeholders in this context include:

- **Public health agencies**
- **Veterinary surveillance teams**

- **Poultry farm operators**
- **Policy-makers responsible for outbreak preparedness**

These stakeholders require rapid, evidence-based tools to help answer questions such as:

- “Should we escalate response in this region?”
- “Is this outbreak likely to grow quickly or remain contained?”
- “How do environmental and contextual conditions impact risk?”

Our recommender system is designed to simulate **micro-moment decision support**, providing quick severity predictions (case_count) based on contextual data like **location, weather, animal type, and disease subtype.** This helps decision-makers act faster and allocate resources more effectively.

While the models developed in this study did not achieve high predictive accuracy, they identified which types of contextual data hold predictive value (e.g., disease type, region) and where gaps exist (e.g., lack of real-time farm movement data). This insight is crucial for guiding future system development, helping stakeholders understand:

- What data needs to be collected
- How it could improve real-time response tools
- Why model choice may not matter as much as data quality

By translating technical analysis into actionable intelligence, this project lays the foundation for better stakeholder-aligned tools in biosecurity and outbreak response systems.

8. Conclusion

The goal of this project was to use machine learning regression models to investigate the predictive capability of contextual and environmental factors in assessing the severity of avian disease outbreaks. We put in place a comprehensive data analytics pipeline, starting with dataset development and exploratory data analysis (EDA) and ending with model selection, evaluation, and interpretation, building on the framework set out in Assessment 1.

We found weak but discernible patterns using EDA between epidemic case counts and environmental variables (such as temperature and rainfall). We decided to test both linear and non-linear models since contextual factors like disease kind and location were found to be more significant. Three models were put into practice and assessed:

Random Forest, Ridge Regression, and Linear Regression. RMSE was used to evaluate each.MAE, and R^2 scores, complemented by visual diagnostics (residual plots, feature importance) and statistical comparison via a paired t-test.

Although all models performed poorly in predictive accuracy ($R^2 < 0$), the analysis provided important insights:

- Data quality, not algorithm complexity, was the primary constraint.
- Random Forest revealed which contextual features (e.g., H5N1 outbreaks, South East region) were more predictive.
- The paired t-test confirmed that model performance differences were not statistically significant, highlighting the need for richer real-time data.

From an organisational standpoint, this study highlights how crucial it is to assist epidemic risk decision-making by integrating epidemiological data with real-time contextual intelligence. The findings highlight the need for more dynamic and detailed data to enable predictive modelling techniques in order for them to be operationally useful for stakeholders, including public health authorities and operators of chicken farms.

The methodology, model evaluation, and stakeholder-aligned interpretation show a strong foundation for iterative development of a robust context-aware recommender system for micro-moment decision support in animal disease surveillance, even though the prototype did not produce strong predictive results.

9. AI Usage Note

We generated synthetic environmental data variables (temperature, humidity, rainfall, and wind speed) with the help of generative AI (ChatGPT). In order to replicate genuine patterns based on widely established weather behaviour, the AI offered advice on suitable statistical distributions. All environmental data was specifically created for the simulation scenario; no real environmental datasets were retrieved or scraped.

ChatGPT provided assistance in organising and summarising this study. The crew carried out and verified every analysis.