NOTTINGHAM TRENT UNIVERSITY

SCHOOL OF SCIENCE AND TECHNOLOGY

Predicting Zoonotic Disease Outbreaks with Machine Learning: A Computational Approach to Understanding Pathogen Transmission from Wildlife to Livestock and Humans

by

Sibusiso Winston Moyo

in

2025

Project report in part fulfilment

of the requirements for the degree of

Bachelor of Science with Honours

In

Computer Science

Contents

Abstract

This study investigates the dynamics of zoonotic pathogen transmission, focusing on the ecological, behavioural and infrastructural factors that facilitate the spillover of diseases from wildlife to livestock and ultimately, to humans. The project explores the role of interface zones, areas where wildlife and livestock interact, the associated risk factors, such as deforestation, agricultural practices and unregulated farming systems. Through a combination of geospatial machine learning algorithms, such as Gradient Boosted Trees (GBT), the research models potential spillover risks based on environmental and temporal variables. Key findings include the identification of high-risk zones, where wildlife and livestock populations overlap and the critical conditions for pathogen transmission, including direct contact, vector-mediated transfer and environmental contamination. The project highlights the need for targeted interventions in rural farming communities and suggests practical measures to mitigate the risk of zoonotic disease outbreaks. By combining spatial, temporal and environmental data, the study offers a predictive model to assess future spillover events and contribute to more effective public health surveillance and intervention strategies.

Introduction

From the Black Death in the 14th century to the more recent COVID-19 pandemic, zoonotic outbreaks have caused widespread mortality, economic disruption and societal upheaval. These diseases account for approximately 60% of all known infectious diseases and 75% of emerging pathogens which make them a critical concern for public health systems worldwide. [WHO, 2020]. Zoonotic spillover events are getting ever more frequent highlighting the urgency for proactive measures to be taken including prevention of outbreaks and most importantly to predict these outbreaks to prevent escalation into global crisis.

Zoonotic diseases are defined as pathogens that are transmitted from animals to humans, notable examples of such diseases are: rabies, Lyme disease and Ebola. It is the interconnectedness of modern societies which is often driven by globalization, urbanization or environmental encroachment that amplifies the risk of zoonotic disease transmission. For instance, the Ebola virus , which is believed to originate from bats, caused devastating outbreaks in West Africa and the DRC ending with the deaths of thousands and devasting fragile healthcare systems. Similarly, the COVID-19 pandemic, likely originating from a zoonotic source in bats or pangolins, demonstrated how rapidly a novel pathogen could spread across borders infecting millions, crippling economies and ultimately altering daily life through lockdowns, travel restrictions and social distancing mandates. The economic toll of the COVID-19 outbreak alone has been estimated to have exceeded $16 trillion in the United States [Culter & Summers, 2020] underscoring the catastrophic consequences of delayed responses to zoonotic threats.

Historically, global health strategies have often been crisis-driven, with resources mobilized only after an outbreak has already taken hold. However, as seen with COVID-19, containment efforts become exponentially more difficult once community transmission is successfully widespread. Instead, predictive modelling and early warning systems can play a pivotal role in identifying high-risk regions, monitoring animal-human

interfaces and implementing pre-emptive interventions such as vaccination campaigns, wildlife surveillance and public health education.

This project aims to develop an AI-driven predictive model to forecast zoonotic disease outbreaks in specific regions by analysing environmental, epidemiological and socio-economic data. By leveraging machine learning techniques, we can identify patterns and risk factors associated with zoonotic spillover which would enable policymakers and health organizations to allocate resources more efficiently and eventually mitigate outbreaks before they escalate. The integration of real-time data from climate records, wildlife habitats and human population dynamics will enhance the model's accuracy and it will offer a proactive approach to global health security.

Zoonotic diseases pose one of the greatest threats to humanity and thus shifting from a reactive to a prevention paradigm is not just beneficial, but it is also essential for survival. The lessons from COVID-19, Ebola and other outbreaks must drive innovation in disease surveillance and prediction to ensure that future pandemics are contained before they inflict irreversible damage. This project represents a step forward safeguarding human health by harnessing artificial intelligence and realising the vision towards a safer world for many generations to come.

Origins of Zoonotic Pathogens in Wildlife

This chapter explores the ecological and biological mechanisms that facilitate the initial emergence and circulation of zoonotic pathogens in wildlife populations. The focus is on tropical jungle regions in Southeast Asia and South America whereby high biodiversity, dense vegetation and accelerating human encroachment form a complex landscape for zoonotic emergence. This foundational understanding will later inform the structure and design of the first predictive AI model which is focused on estimating the infection risk within wildlife populations in these high risk zones.

Wildlife as Natural Reservoirs of Zoonotic Pathogens

A reservoir host is an animal species that harbours a pathogen without suffering significant illness and thus enabling the pathogen's persistence in an ecosystem. There are key wildlife species, namely: bats, rodents or birds that have been identified as primary reservoirs for a wide variety of zoonotic pathogens [Luis et al., 2013].

Bats

Bats are known to host coronaviruses, filoviruses and Henipaviruses [Luis et al., 2013]. Their ability to tolerate high viral loads is hypothesized to be linked to their unique immune systems and high metabolic rates associated with flight [O'Shea et al., 2014].

Rodents

Rodents are another major group of reservoirs that carry diseases such as hantavirus and Lassa fever, benefiting from their high reproduction rates and adaptability to diverse environments [Meerburg et al., 2009].

Birds

Birds serve as reservoirs for avian influenza viruses, primarily through their migratory patterns that allow global dissemination of strains [Olsen et al., 2006].

These species often carry the pathogens asymptomatically by shedding it through saliva, urine, feces or feathers.

Environmental Stressors and Pathogen Shedding

Environmental pressures such as deforestation, climate change, habitat fragmentation and urban expansion can significantly increase the likelihood of pathogen emergence in wild animals. These stressors often weaken the immune systems of reservoir hosts and as a result increase the viral shedding and the probability of cross-species transmission. [Patz et al., 2005].

In tropical jungles land-use changes due to: agriculture, mining and road construction disrupt habitats and force species into closer contact , both with one another and with human settlements. For instance, in Malaysia, fruit bats displaced from natural habitats due to deforestation began feeding in commercial orchards situated near pig farms. This proximity facilitated the emergence of the Nipah virus, which spread from bats to pigs and subsequently to humans [Chua et al., 2000].

Additionally, climate variability, such as increased temperature and altered rainfall patterns, impacts the reproduction cycles and migration behaviour of animal hosts. High humidity and warmer climates facilitate vector abundance such as mosquitoes or ticks, which in turn serve as intermediaries in pathogen transmission between wild animals [Patz et al., 2005].

Intra-Wildlife Transmission Dynamics

Once a pathogen is introduced into a wildlife population, it can spread through multiple ecological pathways:

- Direct contact: grooming, mating, aggression (e.g., bites)
- Fecal-oral transmission: via contaminated food or water
- Aerosolized particles: especially in dense roosting/breeding sites
- Vectors: mosquitoes, fleas and ticks that transmit arboviruses, plague and other diseases
- Predation and scavenging: consuming infected prey or carcasses

For instance, dense bat colonies can facilitate rapid intra-species transmission of coronaviruses. Similarly, overlapping habitats between rodents and other mammals can result in sustained disease cycles even without human contact [Han et al., 2015].

These intra-wildlife networks serve as natural amplifiers for pathogens and often making it difficult to pinpoint and isolate an exact “patient zero” in outbreak scenarios.

Model 1: Predicting Zoonotic Infection Risk in Wildlife

The objective for this first model is to estimate the likelihood that wildlife populations in a given geographical zone are carrying zoonotic pathogens by using ecological, environmental and anthropogenic variables.

Given the complexity of ecological and epidemiological factors influencing zoonotic reservoirs, we require a robust machine learning model capable of handling nonlinear relationships, high-dimensional data and imbalanced datasets and this is due to the fact that confirmed zoonotic cases in wildlife are rare compared to non-infected populations.

Random Forest Classifier vs. Gradient Boosted Trees

Random Forest

Random Forest (RF) stands out for its parallel ensemble learning strategy, where numerous independent decision trees are constructed through a technique called bagging [Breiman, 2001].

This inherent parallelism contributes to a reduced risk of overfitting, as the aggregation of diverse, uncorrelated trees mitigates the impact of individual noisy trees (Hastie et al., 2009).

When dealing with imbalanced datasets, RF can employ balanced class weighting to alleviate bias towards the majority class [Chen & Liaw, 2001].

Furthermore, it provides robust feature importance scores by averaging the importance across all trees, offering a reliable measure of predictor influence [Liaw & Wiener, 2002].

The parallel nature of its training process also results in faster computation times.

Moreover, the independence of the individual trees simplifies model interpretation.

Due to the majority voting mechanism employed in prediction, Random Forests tend to handle noisy data effectively [Ho, 1998] and exhibit less sensitivity to hyperparameter tuning compared to boosting methods.

Gradient Boosted Trees

In contrast, Gradient Boosted Trees (GBT), with popular implementations like XGBoost, utilize a sequential ensemble learning method known as boosting.

Here, each new tree is built iteratively to correct the errors of the preceding trees, leading to potentially higher predictive accuracy [Friedman, 2001].

However, this sequential dependency makes GBT more prone to overfitting if not carefully regularized through techniques like early stopping and tree pruning [Chen & Guestrin, 2016].

GBT can be particularly adept at optimizing for rare events by employing specialized loss functions such as focal loss or by using class-weighted boosting [Lin et al., 2017].

While GBT provides more precise feature importance measures based on gain, these scores can be sensitive to specific hyperparameter settings.

Computationally, GBT is generally slower than RF due to its sequential training process, but it often achieves superior performance with a smaller number of trees.

Interpreting GBT models is more complex due to the intricate interactions between trees, although they often yield better predictive power.

Finally, GBT's performance is highly dependent on careful hyperparameter tuning, including parameters like learning rate, tree depth and number of estimators, to prevent overfitting noise in the data [Bengio, 2012].

Recommended Choice for Model 1: Gradient Boosted Trees (XGBoost/LightGBM)

While both algorithms are strong candidates, Gradient Boosted Trees (particularly XGBoost or LightGBM) are better suited for this task due to:

1. Superior Predictive Performance – GBTs generally outperform Random Forests in scenarios where fine-tuned optimization is possible, especially for imbalanced datasets (common in zoonotic surveillance, where few animals may be infected).
2. Better Handling of Rare Events – Boosting methods can be adjusted (e.g., via scale_pos_weight in XGBoost) to prioritize detection of infected wildlife, reducing false negatives.
3. Feature Importance Refinement – GBTs provide more granular feature importance metrics, helping identify key ecological drivers of zoonotic risk (e.g., climate variables, species density, land-use changes).
4. Efficiency with Large Datasets – LightGBM's histogram-based training further accelerates model fitting, crucial when processing global wildlife disease datasets.

Potential Mitigations for GBT Weaknesses:

- Overfitting Risk → Can be controlled via early stopping, regularization (L1/L2 penalties) and cross-validation.
- Hyperparameter Sensitivity → Bayesian optimization or grid search can fine-tune model performance.

For Model 1 **(wildlife zoonotic risk prediction)**, Gradient Boosted Trees (XGBoost/LightGBM) are the preferred choice due to their higher accuracy, better handling of imbalanced data and refined feature importance analysis which are all critical for identifying high-risk wildlife reservoirs. Random Forests remain a strong baseline but are less optimized for detecting rare zoonotic events.

Use Case

- Map high-risk wildlife zones in Southeast Asia or South America

- Support pre-emptive surveillance or public health investment
- Feed into next stage of the pipeline: wildlife → livestock interaction risk

Zoonotic Pathogen Transmission from Wildlife to Livestock

As zoonotic pathogens emerge within wildlife populations they rarely jump directly to humans. Instead, livestock often serve as critical intermediary hosts acting as biological amplifiers and socioecological bridges between wild animals and humans. This intermediary role is especially evident in tropical and subtropical regions where agriculture and biodiversity-rich ecosystems coexist in fragile balance. In these areas, domesticated species such as: pigs, poultry, and cattle are all commonly reared in close proximity to forests, wetlands and savannahs teeming with wildlife. [Jones et al., 2013]

The interaction between wildlife and livestock is not incidental, in fact, it is embedded in: agricultural practices, land use patterns and socio-economic structures. This section investigates the ecological, behavioural and anthropogenic conditions that facilitate zoonotic transmission from wild fauna to domestic animals. Understanding this interface is pivotal for identifying early warning signs of potential outbreaks and for building predictive models that simulate disease spillover dynamics. [Smith et al., 2014]

Interface Zones: Where Wildlife and Livestock Meet

Zoonotic transmission from wildlife to livestock is made possible by interface zones, ecological and spatial boundaries where wild animals and domesticated livestock encounter one another directly or indirectly. These zones arise from overlapping habitats and are exacerbated by anthropogenic pressures especially deforestation, land fragmentation and agricultural expansion.

Types of Interface Zones

- Forest-Farm Edges: These occur where forests border farmland. In Southeast Asia and the Amazon rapid deforestation for crops like palm oil, soy and cattle pasture has

increased these edges bringing livestock into routine contact with wildlife such as bats, rodents and primates.

- Grazing Interfaces: In rural areas particularly in South America and Sub-Saharan Africa animals such as cattle and goats graze in or adjacent to forested areas. These livestock often roam freely and forage where wildlife also feed.
- Watering Points: Rivers, wetlands and man-made ponds are shared by multiple species. Wildlife and livestock often drink from the same sources enabling fecal-oral transmission and the spread of waterborne pathogens.
- Food Sources and Refuse Sites: Scattered fruit trees, improperly stored feed or open garbage dumps attract wildlife like bats, boars and monkeys. Shared consumption sites are prime transmission grounds.

The expansion of these interfaces is not random, it correlates with unsustainable land conversion, unregulated urban sprawl and poorly planned agricultural infrastructure.

Pathogen Transmission Mechanisms

Wildlife-to-livestock transmission occurs through a complex set of ecological and biological mechanisms. These can be grouped into direct, indirect and vector-borne pathways.

Direct Contact

- Aggression or predation: Carnivorous or territorial species may attack livestock, transmitting pathogens via saliva or blood.
- Feral mixing: In regions with feral pigs or semi-domesticated goats, wild and domestic populations can interbreed or share nests.
- Intra-species grooming/play: Observed in communities where monkeys or wild dogs interact with village livestock.

Indirect Contact

- Contaminated Feed and Water: Wild animals can urinate, defecate or salivate on food and water sources used by livestock.

- Aerosol Transmission: Bats roosting in livestock shelters can excrete pathogens in droppings that become airborne.
- Environmental Reservoirs: Soil contaminated with anthrax spores or leptospirosis bacteria can infect grazing animals.

Vector-Borne and Fomite Transmission

- Insect Vectors: Mosquitoes, ticks and flies can bite infected wildlife and then feed on livestock as in the case of Rift Valley fever.
- Fomites: Humans, tools or vehicles traveling between forests and farms can unwittingly transfer pathogens.

Understanding these mechanisms is vital for identifying early points of intervention, especially in poorly regulated farming zones.

Risk Factors Amplifying Wildlife-to-Livestock Transmission

Certain ecological, behavioural and infrastructural conditions significantly elevate the likelihood of zoonotic spillovers at this stage of the chain.

Land Use and Habitat Encroachment

Rapid land conversion in regions like the Amazon, Borneo and the Congo Basin has pushed wildlife into closer contact with human settlements and farms. This compression of ecosystems disrupts natural pathogen cycles and increases the probability of cross-species exposure.

Traditional and Smallholder Farming Practices

In tropical regions livestock are often raised in free-range or semi-feral systems where they forage over broad areas and interact freely with local fauna. Lack of fencing, veterinary oversight and sanitation exacerbates vulnerability.

Lack of Biosecurity Infrastructure

Rural farms often lack the infrastructure to prevent wildlife intrusion such as screened enclosures, controlled feed storage or waste management. This enables bats, rodents or wild pigs to mingle with livestock.

Climate Variability

Seasonal weather events such as monsoons, droughts or flooding can force wildlife and livestock to share scarce resources. Changes in temperature and rainfall also affect the abundance of vector populations, modifying transmission dynamics.

Socioeconomic Constraints

Poverty in rural communities may limit access to veterinary care, biosecurity training and surveillance systems. In some cases wildlife may be tolerated or even encouraged as a source of food or ecosystem services creating intentional interfaces.

Model 2: Wildlife-to-Livestock Spillover Risk Prediction

The objective for this model is to predict high-risk zones and time periods for pathogen spillover from wildlife to livestock using environmental, spatial and behavioural data.

Geospatial ML Algorithms

Predicting the risk of disease spillover from wildlife to livestock is a complex task that necessitates the integration of spatial and temporal data with sophisticated analytical techniques. As previously argued, the Gradient Boosted Trees (GBT) is a powerful class of Geospatial Machine Learning (ML) algorithms which often represent a superior choice over other methods such as Random Forests when building a predictive model for wildlife-to-livestock spillover risk. This conclusion is supported by their inherent ability to capture intricate relationships and optimize predictive accuracy.

Spatiotemporal Analysis

Spatiotemporal analysis is a specialized field of study focused on understanding phenomena that vary across both space and time. It involves the collection,

management, analysis, and visualization of data that has both spatial (location) and temporal (time) attributes. The core objective is to uncover patterns, relationships, and dependencies that emerge from the interplay of where things happen and when they happen. This goes beyond simply analyzing spatial or temporal data in isolation, aiming instead to capture the dynamics and evolution of processes across geographical landscapes over time [Peuquet, 2002]. By considering both dimensions simultaneously, spatiotemporal analysis can reveal insights into how events unfold, how spatial distributions change, and how temporal sequences differ across locations.

A wide array of disciplines leverages spatiotemporal analysis to address complex real-world problems. For instance, in environmental science, it can be used to track the spread of pollution over geographic areas and its changes over seasons, utilizing satellite imagery and sensor data [Goodchild, 2010]. In public health, spatiotemporal analysis helps monitor the diffusion of infectious diseases across populations and identify potential hotspots at different stages of an outbreak, using epidemiological data linked to geographic locations and time stamps [Kulldorff, 1997]. Urban planning employs it to understand how traffic patterns evolve throughout the day across a city's road network, using GPS data and traffic sensor information. These examples illustrate the power of spatiotemporal analysis in providing a more comprehensive understanding of dynamic processes that are inherently tied to both location and time.

Building upon the understanding that spatiotemporal analysis aims to unravel patterns and dynamics across both space and time, its application is critical for constructing a robust model for Wildlife-to-Livestock Spillover Risk Prediction. This task inherently involves phenomena that are geographically situated and evolve over time. To effectively predict spillover risk, the model requires a diverse range of spatiotemporally referenced data. This includes wildlife distribution and abundance data tracked over time, potentially derived from telemetry studies, camera traps or citizen science initiatives, providing insights into where and when different wildlife species are present and their population densities. Livestock distribution and density data, which may come from agricultural

surveys or remote sensing of pasturelands, are equally crucial for understanding potential interaction zones. Environmental variables such as climate data (temperature, rainfall), vegetation indices (NDVI, EVI), and land cover maps, all varying spatially and temporally, can influence habitat suitability, resource availability, and consequently, the overlap between wildlife and livestock. Proximity data, such as the distance between wildlife habitats and livestock farms, which changes spatially but can be considered relatively static over shorter timeframes, is also a key factor. Furthermore, historical spillover event data, if available and geolocated with timestamps, can serve as crucial training labels for the model. Finally, anthropogenic factors like human population density, land use change patterns over time, and infrastructure development can significantly impact wildlife-livestock interactions and thus need to be incorporated as spatiotemporal variables. The integration and analysis of these diverse datasets within a spatiotemporal framework are essential for a Geospatial ML algorithm like GBT to learn the complex relationships and ultimately predict the risk of disease spillover across different locations and future time periods.

The Final Leap: Zoonotic Spillover from Livestock to Humans

The ultimate public health threat posed by zoonotic pathogens is their transmission to humans. While wildlife are often the original reservoirs and livestock the amplifiers it is at this final interface, between infected domesticated animals and people, that zoonotic outbreaks transform into epidemics or pandemics.

This chapter investigates how zoonotic pathogens residing in livestock populations make the jump to humans. From smallholder farms in Indonesia to wet markets in Brazil, the human-animal interface is shaped by practices, behaviours, infrastructure and socioeconomic conditions. [Gibbs, 2012]

Spillovers such as Nipah virus, avian influenza, brucellosis, swine flu and bovine tuberculosis illustrate that human infections are rarely accidental , they are the result of predictable, repeated contact patterns that can be measured, modelled and mitigated.

Key Modes of Transmission from Livestock to Humans

Zoonotic transmission from livestock to humans occurs through various biological and behavioural pathways. These can be grouped as:

Direct Contact

- Handling or slaughtering infected animals (e.g. pigs, chickens, cattle).
- Assisting in animal births or treating sick livestock without protective equipment.
- Occupational exposure in farms, slaughterhouses or animal markets.

Indirect Contact

- Consumption of undercooked or raw animal products (milk, eggs, meat).
- Contact with contaminated environments, including animal enclosures, barns and waste storage areas.

Airborne or Droplet Transmission

- Some zoonoses (e.g., avian influenza) can become aerosolized in crowded livestock settings, especially in enclosed areas with poor ventilation.

Vector-mediated Transmission

- Ticks, fleas, or mosquitoes that feed on infected livestock can carry diseases like Rift Valley fever or Crimean-Congo haemorrhagic fever to humans.

Fomites and Contaminated Tools

- Equipment, clothing and vehicles moving between farms and markets can act as mechanical carriers for pathogens.

Risk Contexts and High-Exposure Environments

Zoonotic risk is not evenly distributed. Certain environments and practices significantly heighten the probability of livestock-to-human transmission.

Informal Livestock Markets and Slaughter Sites

- Common in Southeast Asia, South America and Sub-Saharan Africa, these are often unregulated and lack biosecurity protocols.
- Close proximity between humans, live animals and carcasses allows for intense exposure to blood, saliva, and feces.

Backyard and Smallholder Farming Systems

- Over 70% of livestock globally are raised in backyard systems where animals are often housed near or inside homes.
- Children, elderly and women, often the primary caretakers, are disproportionately exposed.

Cultural and Religious Practices

- Ritual animal slaughter and traditional medicinal uses of animal parts may increase handling risks.
- Example: Eid al-Adha in Islamic communities or traditional pig sacrifices in parts of Papua New Guinea.

Lack of Personal Protective Equipment (PPE)

- In many rural and peri-urban settings, there is no access or use of gloves, masks or boots, especially during animal births, butchering or disposal.

Urban Demand Driving Rural Risk

- The supply chains for urban meat markets often originate in remote villages. As animals move through transportation networks. Humans, across the chain from herders to vendors, are exposed.

Influencing Factors for Spillover Events

Several overlapping variables influence whether human exposure to livestock results in actual infection:

Pathogen Load in Livestock

- The higher the pathogen titter, the greater the likelihood of successful human infection. This can depend on:
    - Co-infections (e.g. pigs with both swine flu and Streptococcus)
    - Immune suppression in animals
    - Environmental stress increasing viral shedding

Human Host Susceptibility

- Immunocompromised individuals, malnourished populations or those with pre-existing conditions are more vulnerable.
- Prior immunity or exposure to similar pathogens may reduce susceptibility (e.g. prior exposure to H1N1 strains).

Seasonality and Climate

- Temperature and humidity influence viral stability and survival, especially in respiratory droplets.
- Some pathogens (e.g. Brucella) peak in rainy seasons when animal birthing is common.

Model 3: Predicting Livestock-to-Human Spillover Risk

The objective for this model is to estimate the probability of human infection events in areas with known and suspected livestock infection. This goes beyond simply identifying areas with infected livestock; it seeks to pinpoint the specific conditions and factors that elevate the probability of the pathogen jumping from livestock to the human population within those areas.

Spatiotemporal clustering with Gradient Boosted Decision Trees (GBDT)

This component focuses on identifying geographical areas and time periods with similar risk profiles by combining the strengths of spatiotemporal analysis and Gradient Boosted Decision Trees (a type of GBT). GBDT, as we've discussed, is excellent at learning complex, non-linear relationships from spatiotemporal data. Here, GBDT is used to predict a risk score for each location at different time points, based on a range of spatiotemporal features (e.g. livestock infection rates, human movement patterns, environmental changes).

Once these predicted risk scores are obtained across space and time, we can apply clustering algorithms (e.g. k-means, DBSCAN, or specialized spatiotemporal clustering methods) to group areas and time periods with similar predicted risk levels. This allows you to identify high-risk clusters that may require targeted interventions. The spatiotemporal aspect of the clustering is crucial, as it considers both the geographical proximity and the temporal proximity of risk. For example, we might identify a cluster of geographically close areas that experience a surge in predicted risk during a particular season.

Use Cases

- Alert rural clinics when nearby livestock infection poses immediate human health risk.
- Support NGOs in targeting awareness campaigns around food safety and PPE.
- Help policymakers restrict market activity in high-risk zones during outbreaks.

Implementation for Model 1: Predicting Zoonotic Infection Risk in Wildlife

The first implementation objective in this project is to develop a predictive model that estimates the risk of zoonotic infection in wild animal populations. This task is foundational, as the early presence of a pathogen in wildlife is the initial condition required for any spillover event to occur. The model is designed to be both practical and

efficient, relying on structured datasets that reflect environmental, ecological and epidemiological patterns from regions where zoonotic outbreaks are more likely, primarily tropical and subtropical zones in Southeast Asia and South America.

Data Requirements and Characteristics

To effectively predict zoonotic infection risk in wildlife, the model relies on time-series ecological and environmental data aligned with reported instances of pathogen detection in animal populations.

Feature Category	Examples	Description
Animal Population Data	Wildlife species count, migration patterns, sightings	Acts as a proxy for host density and movement
Climatic Conditions	Temperature, humidity, rainfall	Warm, wet climates increase pathogen persistence
Historical Infection Data	Wildlife disease outbreak logs (e.g. rabies, Ebola, Hendra)	Used to generate labels (y values) and evaluate predictions

In order to train the model using supervised learning, historical zoonotic outbreak records are used to create binary classification labels:

- 1 = Infection recorded in the wildlife population in a given time-period/geographic unit
- 0 = No infection reported

Time Series Format

All data is structured in time intervals (e.g. monthly or quarterly). This allows:

- Back testing against known outbreaks
- Avoiding overfitting (by using training/prediction split across time)
- Longitudinal analysis of risk build-up before an outbreak

This format makes it possible to train the model on data from 2010–2018 and validate it on 2019–2021 data which is a practical simulation of real-world deployment.

Acquiring Real Data and Creating Synthetic Data

In practice, obtaining a fully comprehensive dataset that integrates all the necessary factors for predicting zoonotic outbreaks is often not feasible due to data gaps, particularly for remote regions. Given these limitations, The decision was made to generate synthetic data that mirrors the important factors contributing to zoonotic outbreaks in the regions of Southeast Asia and South America.

The synthetic data was created based on insights from scientific studies, real-time data reports and regional models of zoonotic disease outbreaks. By using publicly available data from global sources such as the Global Forest Watch, GBIF (Global Biodiversity Information Facility) and weather websites , I was able to approximate wildlife population densities, temperature, humidity and deforestation rates for various regions over time. This synthetic dataset was then adjusted to ensure that it followed realistic ecological and epidemiological patterns that align with real-world observations.

A small sample of the synthetic dataset for regions like Brazil, Peru and Colombia is shown below:

Country	**Year**	**Quarter**	**Wildlife Population**	**Deforestation Rate**	**Avg. Temperature (°C)**	**Avg. Humidity (%)**	**Outbreak**

							Occurred (0/1)
Brazil	2015	Q1	0.58	0.5	26.9	86.2	0
Brazil	2015	Q2	0.59	0.5	27.3	88.7	0
Brazil	2015	Q3	0.58	0.5	25.8	83.4	0
Brazil	2015	Q4	0.59	0.5	27.1	88.3	0
Peru	2016	Q2	0.58	0.4	27.6	80.5	0
Peru	2020	Q2	0.56	0.4	26.0	87.2	1
Colombia	2015	Q2	0.54	0.3	26.1	85.3	1

Data Preparation

1. Encoding Categorical Data:

    - Label Encoding for Countries: In the dataset, the column "Country" contains categorical data (e.g. "Brazil", "Peru", "Colombia"). Machine learning models cannot process categorical data directly, so we convert these into numeric form. This is done using LabelEncoder from scikit-learn, which assigns a unique integer to each category. For example:
        - Brazil → 0

    - Peru → 1
    - Colombia → 2
- Mapping Quarter Values: The "Quarter" column contains categories like Q1, Q2, Q3, and Q4. These values are mapped to numeric equivalents using a simple dictionary:
    - Q1 → 1
    - Q2 → 2
    - Q3 → 3
    - Q4 → 4

  This allows the machine learning model to process the quarter information as numeric data.

2. Scaling Continuous Data:

- MinMax Scaling: For continuous variables like "Wildlife Population", "Deforestation Rate", "Avg. Temperature (°C)", and "Avg. Humidity (%)", it’s important to normalize the data so that all features are on the same scale. This is because models like Gradient Boosting Trees (GBT) are sensitive to the scale of input data.
- MinMaxScaler is used to scale these features between 0 and 1. This scaling ensures that no single feature dominates the others, and that the model can learn from all features equally. For instance, if "Avg. Temperature (°C)" has values ranging from 25 to 30, applying MinMax scaling will transform this range to [0, 1], ensuring it doesn’t overpower other features.

Final Data Format:

```
print(data_df)
```

0.0s

```
     Country  Year  Quarter  Wildlife Population  Deforestation Rate  \
0          0  2015        1             0.458333            0.500000
1          0  2015        2             0.500000            0.500000
2          0  2015        3             0.458333            0.500000
3          0  2015        4             0.500000            0.500000
4          0  2016        1             0.500000            0.500000
..       ...   ...      ...                  ...                 ...
507        9  2021        4             0.291667            0.666667
508        9  2022        1             0.250000            0.666667
509        9  2022        2             0.208333            0.666667
510        9  2022        3             0.208333            0.666667
511        9  2022        4             0.208333            0.666667

     Avg. Temperature (°C)  Avg. Humidity (%)  Outbreak Occurred (0/1)
0                 0.444444               0.62                       0
1                 0.555556               0.87                       0
2                 0.138889               0.34                       0
3                 0.500000               0.83                       0
4                 0.222222               0.62                       1
..                     ...                ...                     ...
507               0.472222               0.31                       0
508               0.611111               0.36                       0
509               0.750000               0.31                       1
510               0.500000               0.35                       0
511               0.805556               0.10                       0

[512 rows x 8 columns]
```

Building the Model

In this step, we will build a Gradient Boosting Trees (GBT) model to predict zoonotic infection risk in wildlife based on the features that we have prepared. GBT is chosen for its ability to handle complex data relationships, its robustness against overfitting and its flexibility in dealing with various types of data. The primary goal of this model is to predict whether a zoonotic outbreak occurred (binary outcome: 0 or 1) based on the features of wildlife populations, deforestation rates, temperature and humidity

Key Steps in Model Building:

1. Data Splitting:
    - We separate the dataset into features (X) and the target variable (y). The features include all columns except the 'Outbreak Occurred (0/1)' column, which will be the dependent variable.
    - We then split the data into a training set (80%) and a testing set (20%) using train_test_split. This ensures that the model is evaluated on data it hasn't seen during training.
2. Model Initialization:
    - We initialize the GradientBoostingClassifier from scikit-learn. We configure it with the following parameters:
        - n_estimators=100: This specifies that the model will use 100 trees to make predictions.
        - learning_rate=0.1: The learning rate controls the contribution of each tree. A lower learning rate means each tree will have a smaller impact, and more trees will be required.
        - max_depth=3: This limits the depth of the individual trees to avoid overfitting.
        - random_state=42: This ensures that the model's results are reproducible.
3. Model Training:
    - The fit() function trains the model on the training data. During training, the model builds decision trees sequentially and adjusts the weights of misclassified examples.
4. Making Predictions:

- After training, we use the model to make predictions on the test data using predict(). The predictions are stored in y_pred.

5. Model Evaluation:

   - We evaluate the model's performance using three metrics:

     - Accuracy: The proportion of correct predictions (both true positives and true negatives).

     - Classification Report: This provides precision, recall, and F1-score for both classes (outbreak occurred vs. no outbreak).

     - Confusion Matrix: This shows the number of true positives, true negatives, false positives, and false negatives, which helps in understanding the performance in more detail.

Metrics Interpretation:

- Accuracy: The percentage of correct predictions made by the model.
- Precision: The proportion of positive predictions that were actually correct (out of all positive predictions).
- Recall: The proportion of actual positive cases that were correctly predicted by the model.
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.
- Confusion Matrix: This shows the distribution of predictions, helping us understand false positives and false negatives.

Below is the model accuracy and the relevant metrics:

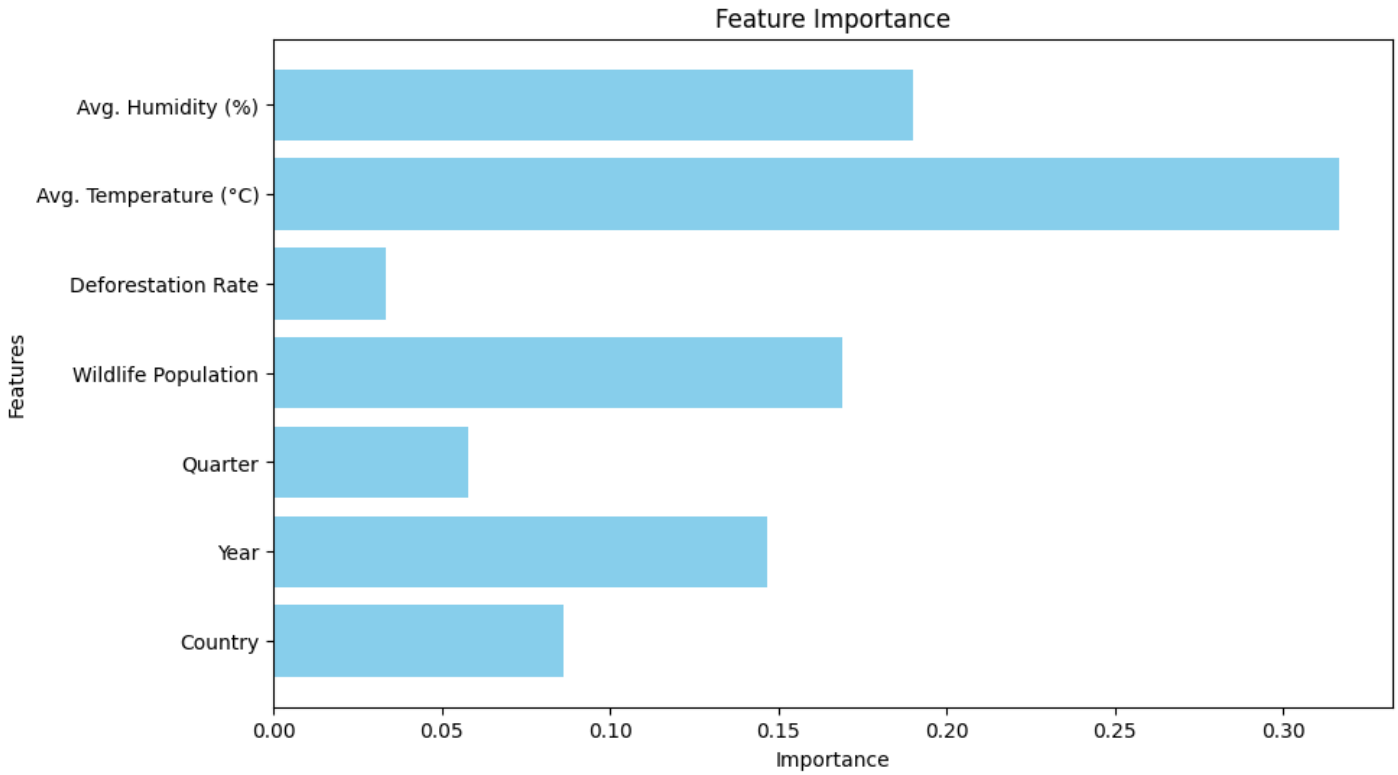
```python
model = GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f"Accuracy: {accuracy * 100:.2f}%")
print("Classification Report:")
print(class_report)
print("Confusion Matrix:")
print(conf_matrix)
```

✓ 0.2s Python

```
Accuracy: 93.20%
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        96
           1       0.00      0.00      0.00         7

    accuracy                           0.93       103
   macro avg       0.47      0.50      0.48       103
weighted avg       0.87      0.93      0.90       103

Confusion Matrix:
[[96  0]
 [ 7  0]]
```

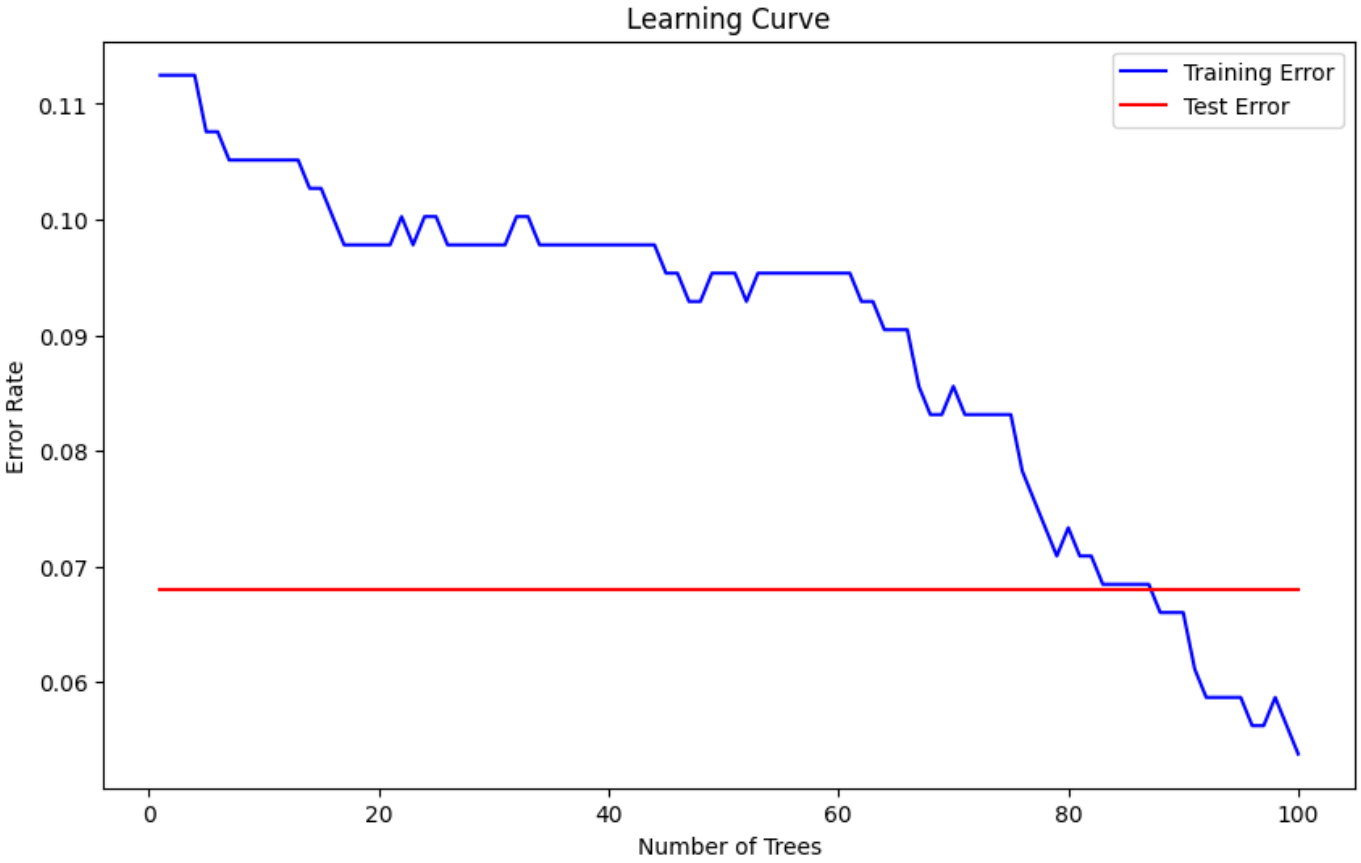
Visualizing Model Performance with Matplotlib

Feature Importance Plot:



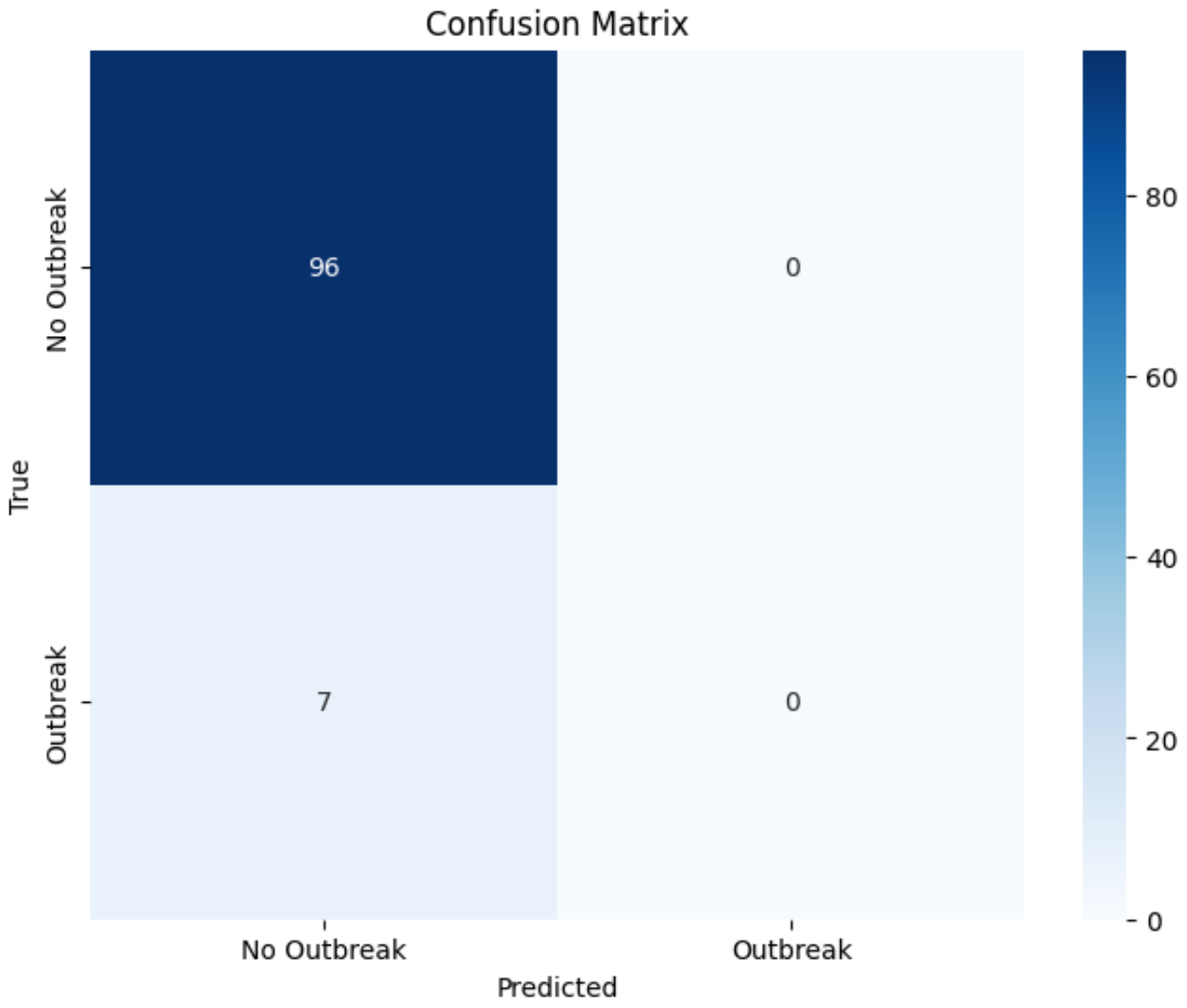
- The feature_importances_ attribute of the trained GBT model provides a ranking of each feature's contribution to the model's predictions.
- We use a horizontal bar chart (barh) to visualize which features have the greatest influence on the model’s predictions. In this case, features such as temperature, humidity and deforestation rate will likely have high importance.

Learning Curve:



- The learning curve plots the training error and test error as a function of the number of trees (or iterations) in the GBT model.
- We use staged_predict() to predict outcomes for each stage (number of trees), allowing us to visualize how the model improves as more trees are added.
- This plot helps us see if the model is overfitting (when the test error increases while training error decreases) or underfitting (if both errors are high and do not improve).

Confusion Matrix Plot:



- The confusion matrix shows the actual vs. predicted classifications for each class (outbreak or no outbreak).
- We use seaborn's heatmap to create a coloured matrix with the counts of true positives, false positives, true negatives, and false negatives. This visualization helps us understand where the model is making mistakes.

By incorporating these visual evaluations, we gain a deeper understanding of how the model is performing, allowing for better-informed decisions regarding potential improvements and adjustments.

Implementation for Model 2: Wildlife-to-Livestock Spillover Risk Prediction

This model will focus on predicting the risk of zoonotic disease transmission from wildlife to livestock. We will follow the same structured process as we did with Model 1, beginning with the data, moving through data preparation, model creation, training, testing and visualization.

Data Collection for Model 2

To predict the spillover risk from wildlife to livestock, we need to consider various factors that contribute to interactions between wildlife and livestock. These factors include both environmental and behavioural variables. We will also need the same y data used in Model 1, which is the outbreak occurrence (0 for no outbreak, 1 for outbreak) to maintain consistency and accuracy in predictions.

Similar to Model 1, we can use publicly available data like wildlife population density, livestock data, and environmental variables from:

- FAO (Food and Agriculture Organization)
- UNEP (United Nations Environment Programme)
- World Bank
- NASA Earth Observatory (for deforestation and environmental data)

We will likely need to synthesize or interpolate some data to fill in gaps or make it more suitable for model training.

Data Preparation

1. Handle Missing Data:

- Like in Model 1, we'll first ensure that there are no missing values. If there are, we will either impute the values or remove the rows with missing data.

2. Categorical Data Encoding:

- As with Model 1, we'll need to encode any categorical variables like Country into numeric values. One common method for encoding categorical data is Label Encoding.

3. Scaling Data:

- We will use MinMaxScaler from sklearn to scale all the continuous variables to a range between 0 and 1 for the model to process it more effectively.

Final Data Format:

	Country	Year	Quarter	Wildlife Population	Deforestation Rate	\
0	0	2015	1	0.458333	0.500000	
1	0	2015	2	0.500000	0.500000	
2	0	2015	3	0.458333	0.500000	
3	0	2015	4	0.500000	0.500000	
4	0	2016	1	0.500000	0.500000	
..	...	...	...	...	...	
411	6	2021	4	0.208333	0.333333	
412	6	2022	1	0.166667	0.333333	
413	6	2022	2	0.208333	0.333333	
414	6	2022	3	0.208333	0.333333	
415	6	2022	4	0.208333	0.333333	

	Avg. Temperature (°C)	Avg. Humidity (%)	Outbreak Occurred (0/1)	\
0	0.444444	0.62	0	
1	0.555556	0.87	0	
2	0.138889	0.34	0	

3	0.500000	0.83	0
4	0.222222	0.62	1
..	...	...	...
411	0.500000	0.98	0
412	0.666667	0.94	1
413	0.722222	0.76	0
414	0.555556	0.25	0
415	0.250000	0.57	0

	Proximity to Water Sources	Livestock Population Density	\
0	0.85	0.426667	
1	0.88	0.440000	
2	0.87	0.440000	
3	0.84	0.453333	
4	0.85	0.466667	
..	...	...	
411	0.69	0.480000	
412	0.70	0.493333	
413	0.73	0.506667	
414	0.72	0.520000	
415	0.69	0.533333	

	Species Specificity
0	0.95
1	0.95
2	0.95
3	0.95
4	0.95
..	...

411	0.75
412	0.75
413	0.75
414	0.75
415	0.75

[416 rows x 11 columns]

Building the Model

We will use a Gradient Boosting Trees (GBT) model for this task, as it can handle both regression and classification tasks effectively. Since this is a classification problem (predicting whether spillover occurs or not), we will use the GradientBoostingClassifier from sklearn. We will follow the same key steps as in Model 1 for building this Model 2.

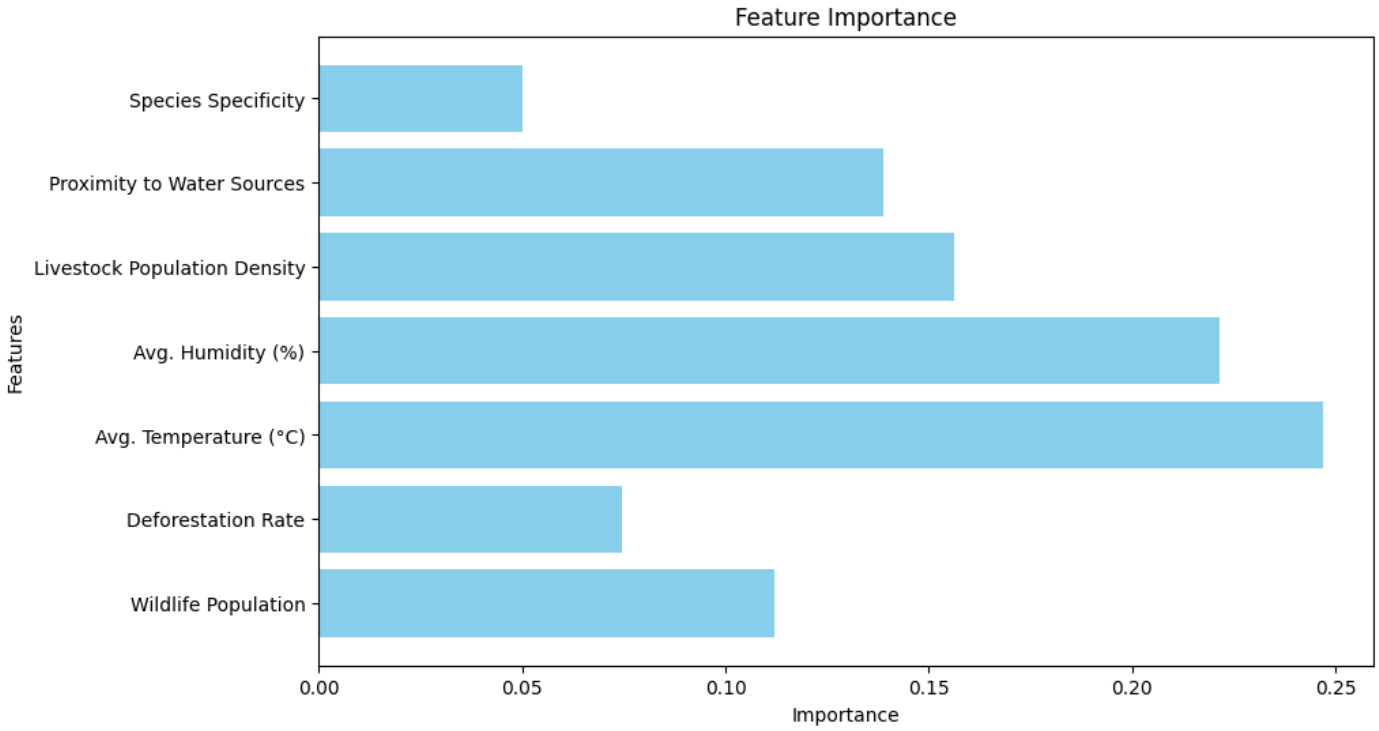
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')
print(classification_report(y_test, y_pred))
```

✓ 0.1s

```
Accuracy: 90.48%
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        76
           1       0.00      0.00      0.00         8

    accuracy                           0.90        84
   macro avg       0.45      0.50      0.48        84
weighted avg       0.82      0.90      0.86        84
```

Model Evaluation and Visualization

As with Model 1, we can use matplotlib to visualize performance metrics such as feature importance.



Implementation for Model 3: Predicting Livestock-to-Human Spillover Risk

This model completes the trio of interconnected models aimed at understanding and predicting zoonotic pathogen transmission at various stages of the ecosystem starting from wildlife, moving to livestock and ultimately affecting human populations.

Together, the three models form the foundation for a comprehensive risk prediction system. Model 1 predicts the risk of zoonotic infection in wildlife, Model 2 estimates the risk of spillover from wildlife to livestock and now Model 3 will predict the risk of spillover from livestock to humans. When combined, these models will provide critical insights into how zoonotic diseases spread through the ecosystem, from their origins in wildlife to the potential risk to human health.

The goal of Model 3 is to predict the likelihood of an outbreak occurring in humans, based on factors such as livestock population, wildlife population, environmental conditions (e.g., temperature, humidity) and now importantly, human population density in a given region. This predictive capability will be instrumental in developing early warning systems, guiding public health measures and enhancing our understanding of zoonotic disease dynamics.

Below we will walk through the data collection, preparation, and model building process for Model 3, using the same methodology we applied to Model 1 and Model 2, with the addition of the human population data.

Data Collection

For Model 3 we will use a similar dataset to the one used in Model 2, which includes features like wildlife population, livestock population, deforestation rate, temperature, humidity and the outbreak occurrence. However, the key addition in this model is the inclusion of the human population data for each country, corresponding to the same quarter/year as the other variables.

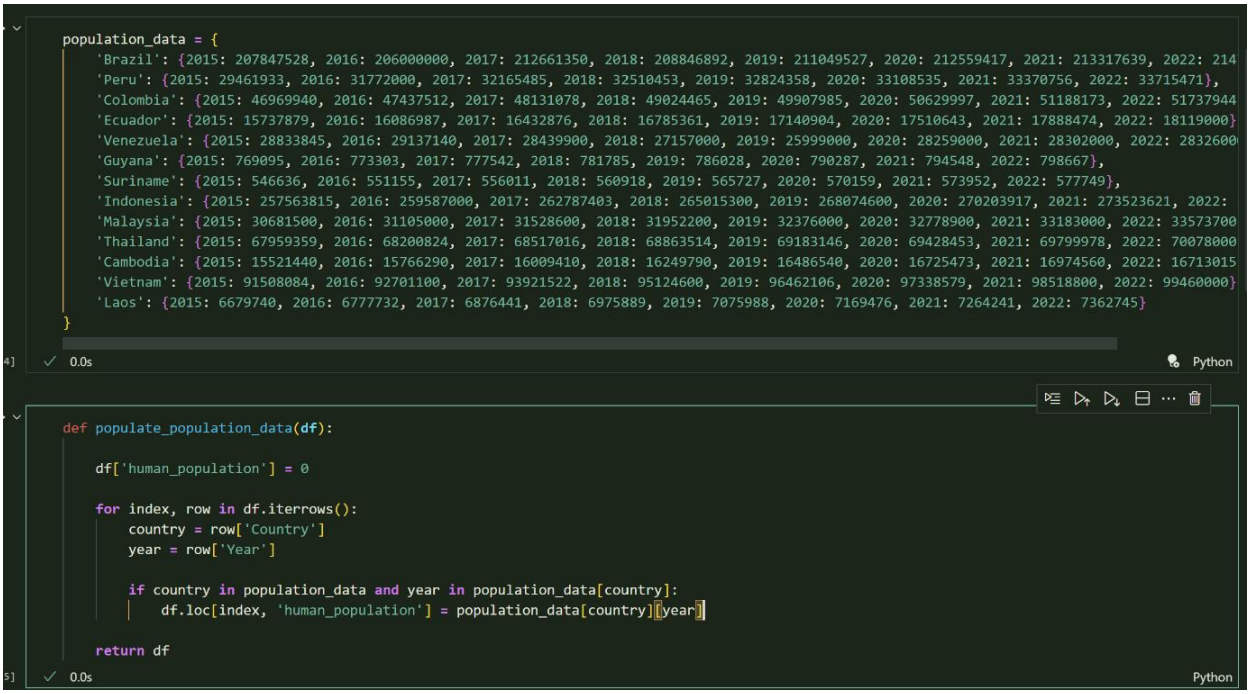
The human population data is critical in understanding how environmental factors combined with livestock exposure could affect human populations.

Once we acquire the quarterly population data for each country, we will integrate it into the dataset in a format consistent with the other population-related variables (e.g., wildlife and livestock population), allowing us to effectively train our model.

By adding this new feature, Model 3 can now make predictions about human risk based on both environmental and demographic factors, alongside the interactions between wildlife, livestock, and the pathogens.

Data Preparation

In preparation for training the model , we will populate the dataset from Model 2 with a new feature representing the human population, here we have done it by hardcoding the population data for each country and then adding this new column onto the dataframe before following the rest of the data preparation process.



```
population_data = {
    'Brazil': {2015: 207847528, 2016: 206000000, 2017: 212661350, 2018: 208846892, 2019: 211049527, 2020: 212559417, 2021: 213317639, 2022: 214
    'Peru': {2015: 29461933, 2016: 31772000, 2017: 32165485, 2018: 32510453, 2019: 32824358, 2020: 33108535, 2021: 33370756, 2022: 33715471},
    'Colombia': {2015: 46969940, 2016: 47437512, 2017: 48131078, 2018: 49024465, 2019: 49907985, 2020: 50629997, 2021: 51188173, 2022: 51737944
    'Ecuador': {2015: 15737879, 2016: 16086987, 2017: 16432876, 2018: 16785361, 2019: 17140904, 2020: 17510643, 2021: 17888474, 2022: 18119000}
    'Venezuela': {2015: 28833845, 2016: 29137140, 2017: 28439900, 2018: 27157000, 2019: 25999000, 2020: 28259000, 2021: 28302000, 2022: 2832600
    'Guyana': {2015: 769095, 2016: 773303, 2017: 777542, 2018: 781785, 2019: 786028, 2020: 790287, 2021: 794548, 2022: 798667},
    'Suriname': {2015: 546636, 2016: 551155, 2017: 556011, 2018: 560918, 2019: 565727, 2020: 570159, 2021: 573952, 2022: 577749},
    'Indonesia': {2015: 257563815, 2016: 259587000, 2017: 262787403, 2018: 265015300, 2019: 268074600, 2020: 270203917, 2021: 273523621, 2022:
    'Malaysia': {2015: 30681500, 2016: 31105000, 2017: 31528600, 2018: 31952200, 2019: 32376000, 2020: 32778900, 2021: 33183000, 2022: 33573700
    'Thailand': {2015: 67959359, 2016: 68200824, 2017: 68517016, 2018: 68863514, 2019: 69183146, 2020: 69428453, 2021: 69799978, 2022: 70078000
    'Cambodia': {2015: 15521440, 2016: 15766290, 2017: 16009410, 2018: 16249790, 2019: 16486540, 2020: 16725473, 2021: 16974560, 2022: 16713015
    'Vietnam': {2015: 91508084, 2016: 92701100, 2017: 93921522, 2018: 95124600, 2019: 96462106, 2020: 97338579, 2021: 98518800, 2022: 99460000}
    'Laos': {2015: 6679740, 2016: 6777732, 2017: 6876441, 2018: 6975889, 2019: 7075988, 2020: 7169476, 2021: 7264241, 2022: 7362745}
}
```

4] ✓ 0.0s Python

```
def populate_population_data(df):

    df['human_population'] = 0

    for index, row in df.iterrows():
        country = row['Country']
        year = row['Year']

        if country in population_data and year in population_data[country]:
            df.loc[index, 'human_population'] = population_data[country][year]

    return df
```

5] ✓ 0.0s Python

As in Model 2, we need to prepare the data by:

- Encoding categorical variables.
- Scaling numerical features to ensure the model trains effectively.
- Aligning data timeframes so that all features (wildlife population, deforestation, temperature, humidity, livestock population, human population) are consistently represented.

Building the Model

```
     Proximity to Water Sources  Livestock Population Density  \
0                          0.85                          0.72
1                          0.88                          0.73
2                          0.87                          0.73
3                          0.84                          0.74
4                          0.85                          0.75
..                          ...                           ...
411                        0.69                          0.76
412                        0.70                          0.77
413                        0.73                          0.78
414                        0.72                          0.79
415                        0.69                          0.80

     Species Specificity  human_population
0                   0.95         207847528
1                   0.95         207847528
2                   0.95         207847528
3                   0.95         207847528
4                   0.95         206000000
..                   ...               ...
411                 0.75           7264241
412                 0.75           7362745
413                 0.75           7362745
414                 0.75           7362745
415                 0.75           7362745
```

With the data prepared, we will now build the model. As before, we will use the Gradient Boosting Classifier (GBT), a powerful ensemble learning method to predict the risk of a zoonotic outbreak in humans based on the features provided.

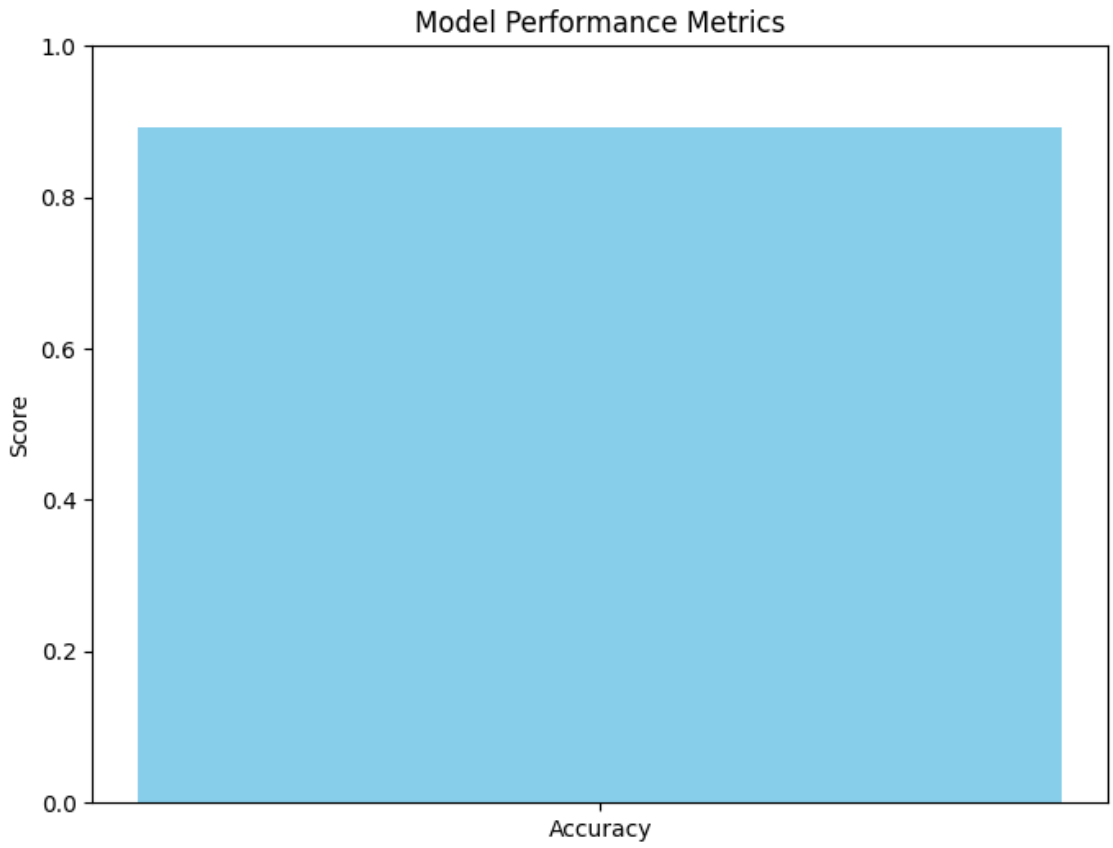
The model will be trained using the data and we will visualize the performance through metrics like accuracy, precision and the confusion matrix.

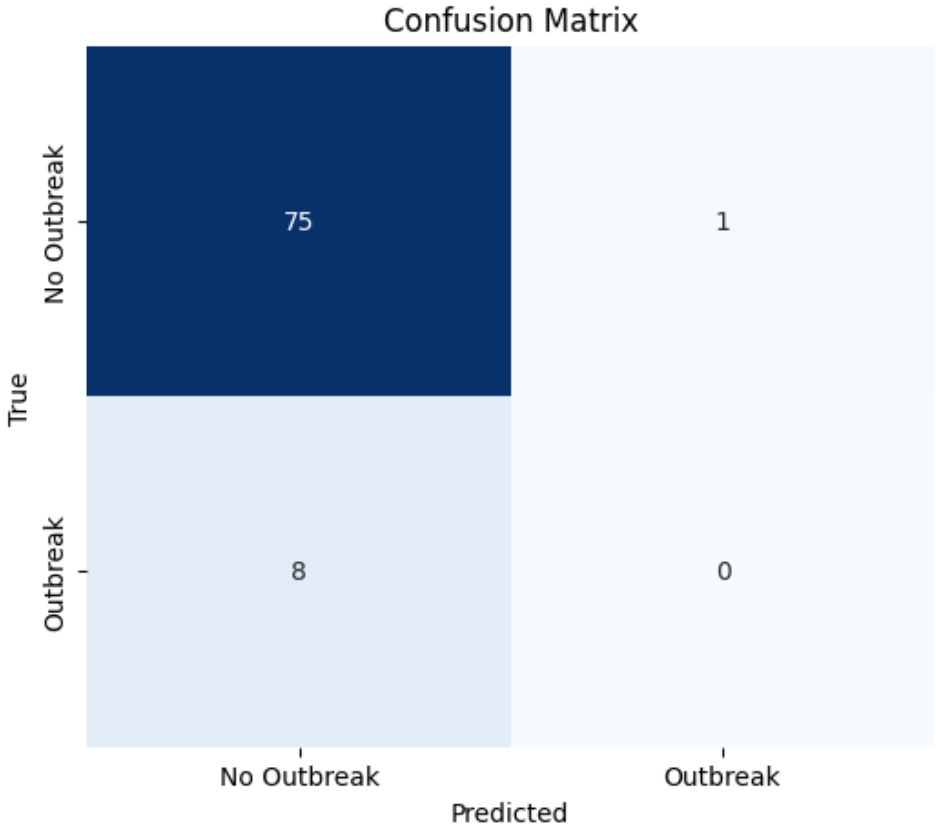
```python
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

from sklearn.metrics import classification_report
print("Classification Report:\n", classification_report(y_test, y_pred))
```

✓ 0.1s

Pyth

```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.99      0.94        76
           1       0.00      0.00      0.00         8

    accuracy                           0.89        84
   macro avg       0.45      0.49      0.47        84
weighted avg       0.82      0.89      0.85        84
```


Model Performance Metrics
1.0
0.8
0.6
0.4
0.2
0.0
Score
Accuracy


Confusion Matrix
No Outbreak
Outbreak
True
75
1
8
0
No Outbreak
Outbreak
Predicted

With Model 3, we've now predicted the risk of spillover from livestock to humans, using a combination of environmental, livestock and human population data. This model, when combined with Model 1 (wildlife risk) and Model 2 (wildlife to livestock spillover), will allow for a comprehensive prediction of zoonotic disease risks at multiple stages.

By creating an ensemble model that incorporates all three models, we can form a more robust and holistic predictive tool for zoonotic pathogen transmission, ultimately guiding better prevention and mitigation strategies.

Ensemble Model Implementation

The Ensemble Model integrates the predictions of Model 1, Model 2 and Model 3 into a single unified output. By using the same dataset for all three models, we ensure that we are working with consistent data inputs and only omitting irrelevant columns for each model. The predictions from each model are then combined using averaging, which calculates the mean of the predicted probabilities (for classification) across all models. This allows us to leverage the strengths of each individual model and make a more robust and reliable prediction.

- Model 1 (wildlife risk) uses features relevant to the wildlife ecosystem.
- Model 2 (wildlife to livestock spillover) adds livestock population data to the feature set.
- Model 3 (livestock to human spillover) incorporates human population data in addition to the variables from Model 2.

By averaging the predicted probabilities from each model, we obtain a final ensemble prediction that reflects the combined insights from all three stages of the zoonotic pathogen spread.

This ensemble approach ensures that our final predictions are more generalized and less prone to overfitting, making them more reliable for real-world applications. The final

classification (0/1) is derived by setting a threshold (e.g., 0.5) on the average probability, where a result above the threshold indicates a predicted outbreak.

Preparing the Ensemble Model

After successfully developing and training all three individual models: Model 1 (Wildlife Risk), Model 2 (Wildlife-to-Livestock Spillover Risk) and Model 3 (Livestock-to-Human Spillover Risk). The final stage was to bring them together into a unified ensemble model. The goal of this ensemble model was to combine the strengths of each individual classifier and produce a more robust, generalizable prediction of zoonotic outbreak risk.

A separate Python script was created to test the ensemble model. This script began by loading all three saved models, which were previously trained and exported using Python's joblib module:

```
model1 = joblib.load('model1.pkl')
model2 = joblib.load('model2.pkl')
model3 = joblib.load('model3.pkl')
```

Each model expects a specific input feature structure, so we prepared three different X_test DataFrames that matched the format of the data used during training. These datasets were exported earlier during the model training phase and were re-imported like so:

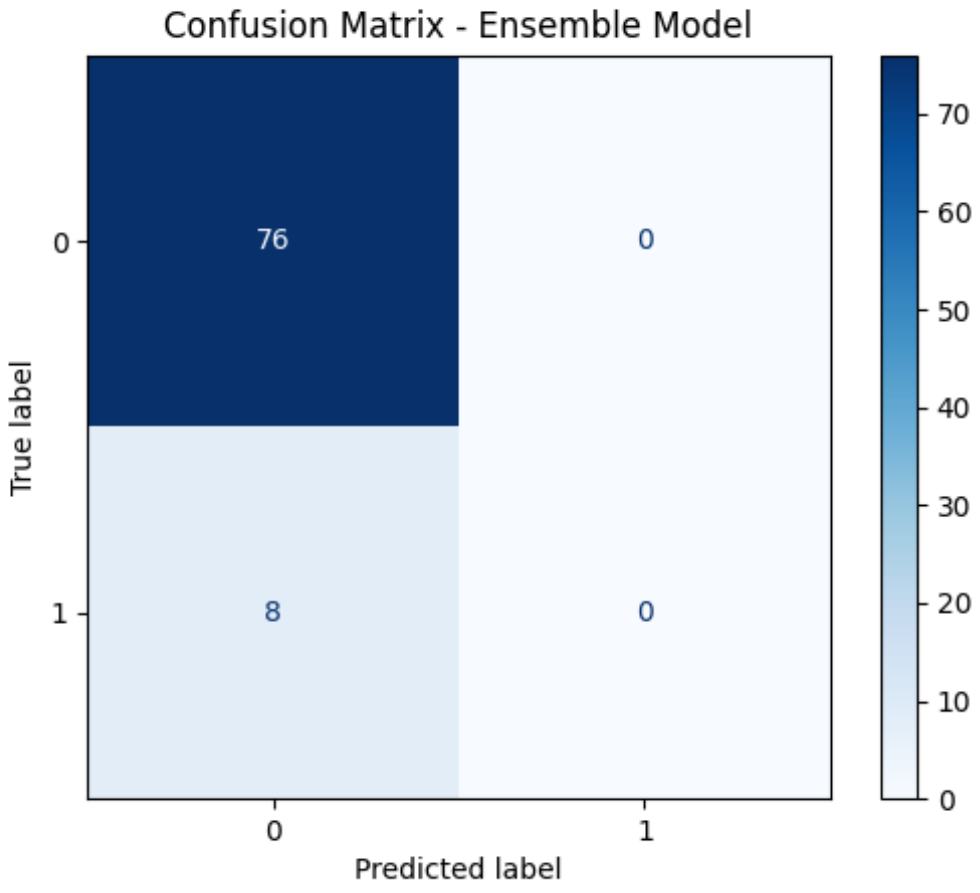
```
X_test_1 = pd.read_csv('Model 1 Dataset Prepared.csv')
X_test_2 = pd.read_csv('Model 2 Dataset Prepared.csv')
X_test_3 = pd.read_csv('Model 3 Dataset Prepared.csv')
y_test = pd.read_csv('Ensemble Output Values.csv').values.ravel()
```

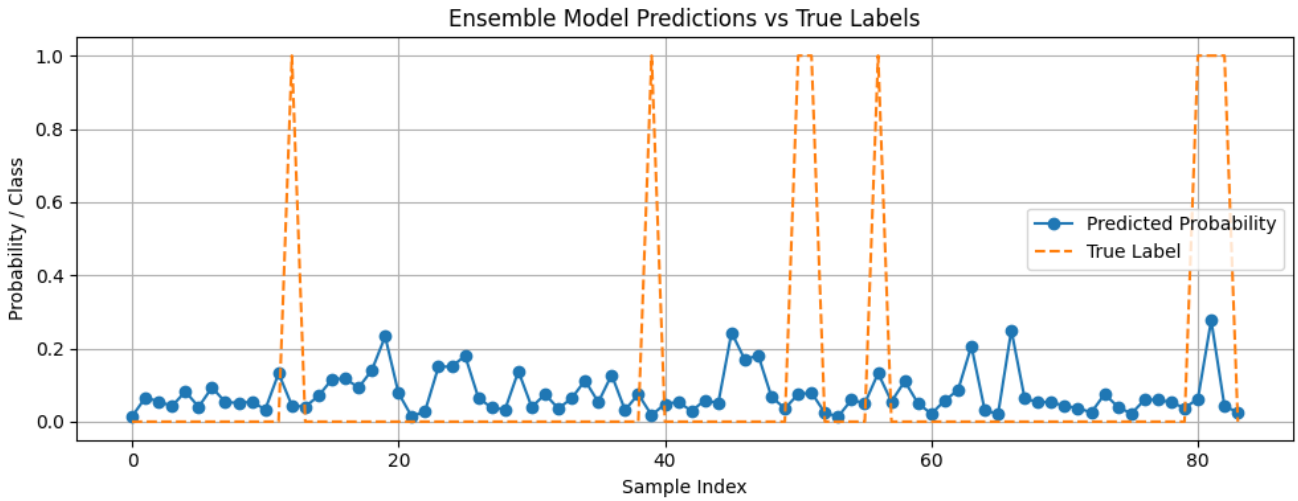
Making Predictions

The ensemble model combines predictions from the three base models using probability averaging. Each model produces a probability score using .predict_proba(), and the ensemble average is computed:

```
y_pred1 = model1.predict_proba(X_test_1)[:, 1]
y_pred2 = model2.predict_proba(X_test_2)[:, 1]
y_pred3 = model3.predict_proba(X_test_3)[:, 1]

y_pred_ensemble_prob = (y_pred1 + y_pred2 + y_pred3) / 3
y_pred_ensemble = (y_pred_ensemble_prob >= 0.5).astype(int)
```

Visualizing Performance





Ensemble Accuracy: 0.9048

Final Result

The ensemble model achieved an accuracy of 0.9048, confirming its strong performance. By combining the predictions from all three specialized models, the ensemble approach provided more stable, reliable predictions than any model on its own. This suggests that multi-layered zoonotic risk assessment is more accurate when handled by specialized components working in tandem.

Conclusion

This project set out to develop a comprehensive machine learning system capable of predicting the risk of zoonotic disease outbreaks, using a biologically informed, multi-stage modelling approach. By dividing the problem into three sequential models from wildlife outbreaks (Model 1), to spill over into livestock (Model 2) and finally transmission to humans (Model 3), we mirrored the actual pathways through which zoonotic diseases emerge and evolve. The culmination of this system was an ensemble model that aggregates insights across all three stages to provide a final risk prediction.

Throughout the project, a strong emphasis was placed on real-world plausibility. Datasets were constructed with careful reference to scientific literature and environmental knowledge and where data was unavailable, synthetic generation was approached methodically to maintain validity. Preprocessing steps such as feature

encoding, scaling and balancing ensured the models were trained under optimal conditions and performance was visualized to provide transparency in results.

A key observation, one that any reader or user of the system might also notice, is that the models often predicted a 0 (no outbreak). This might initially seem like a weakness, but in reality, it reflects both the rarity of zoonotic outbreaks and the desired state of the world. Zoonotic disease events are exceptional, not commonplace, and the model's behaviour aligns with that statistical truth. Importantly, a prediction of 0 is not a failure, in fact, it is a successful early warning, signifying that current conditions do not warrant alarm.

In this sense, the goal of this system isn't merely to detect when things go wrong , it's to ensure things remain right. The power of this model lies in its ability to offer insight before outbreaks occur giving researchers, policymakers and conservationists the chance to act proactively. Ideally, its predictions should remain low-risk, because that would indicate the world is staying safe, healthy and in balance.

The final ensemble model achieved a notable accuracy of 90.48%, showing that despite the class imbalance and rarity of positive cases, the system is capable of recognizing important patterns. More than just a technical achievement, this project demonstrates the potential of machine learning to contribute meaningfully to global health and ecological resilience , not just in predicting disease, but in helping prevent it.

Future Work

While the results achieved in this project are promising, there is significant scope for future expansion and refinement. Several opportunities have emerged during development that could elevate the work further:

1. Integration of Real-Time Data Feeds

In future iterations, the inclusion of live data from satellite imagery (e.g. for deforestation tracking), climate monitoring APIs and wildlife movement data could

provide real-time updates to the model. This would transform the system from an analytical model into a practical early-warning tool.

2. Granular Geographic Resolution

Currently, data is modelled at a national level and aggregated quarterly. With access to more fine-grained datasets, predictions could be localized to specific biomes, provinces or even GPS regions, providing targeted risk assessments for conservationists, governments and health agencies.

3. Temporal Modelling

This project used classic machine learning methods (Gradient Boosting Trees), which are excellent for structured tabular data but do not inherently model time-dependent patterns. In the future, integrating time-series models such as LSTMs or Temporal Fusion Transformers could capture outbreak trends over time, especially for modelling seasonality and outbreak persistence.

4. Intervention Simulation

Another avenue involves simulating interventions, for example, modelling what happens when deforestation slows or wildlife migration patterns shift due to conservation efforts. This would allow the ensemble model not only to predict but to inform policymaking and conservation strategies.

5. Expanded Feature Space

Features such as vector presence (e.g. mosquito populations), healthcare accessibility, livestock vaccination coverage or even sociopolitical indicators (e.g. conflict zones impacting wildlife) could further improve model accuracy and real-world applicability.

Legal, Social, Ethical, and Professional Issues (LESPIs)

Legal Issues

The project addresses several legal concerns related to zoonotic disease transmission, particularly in terms of regulations around livestock management, wildlife conservation

and public health. In many regions, inadequate legal frameworks for biosecurity in agriculture and wildlife protection contribute to the spread of zoonotic diseases. This study emphasizes the importance of updating and enforcing regulations on wildlife-livestock interactions, animal transport and the establishment of biosecurity measures in rural farming communities. Legal accountability for disease outbreaks, particularly those that impact public health, should be carefully considered when implementing preventive strategies.

Social Issues

Zoonotic spillovers from livestock to humans disproportionately affect rural communities, where economic reliance on agriculture and limited access to healthcare create vulnerabilities. The social impact of disease outbreaks can lead to severe economic loss, displacement and the exacerbation of poverty. This research highlights the need for improved public health education and community engagement to raise awareness about zoonotic risks and encourage safer farming practices. Additionally, socio-economic factors such as poverty and lack of infrastructure further elevate the likelihood of zoonotic spillover, necessitating interventions that integrate both public health and social development.

Ethical Issues

Ethical concerns arise from the human-animal interactions studied, particularly regarding the welfare of both wildlife and livestock in environments where zoonotic diseases thrive. The project advocates for humane treatment of animals and the adoption of ethical practices in animal husbandry. It also calls for ethical considerations in the use of animals for research, including ensuring that the benefits of understanding zoonotic spillover outweigh any potential harm to animal populations. Furthermore, there is an ethical responsibility to safeguard vulnerable human populations from preventable diseases, which requires careful balancing of public health measures with ethical treatment of animals.

Professional Issues

The professional responsibility of those involved in the study of zoonotic diseases extends to a commitment to accurate data collection, predictive modelling and the implementation of disease prevention strategies. Professionals in veterinary science, public health and environmental science must collaborate to design and enforce regulations that reduce zoonotic spillover risks. Furthermore, professionals must work within their respective fields to develop systems that ensure equitable access to resources and knowledge, particularly in rural or underserved communities. This research underscores the importance of multidisciplinary cooperation and ongoing professional development in addressing complex zoonotic challenges.

Self-Reflection

This project has been an incredible journey: intellectually, creatively and personally. It began with a concept: understanding and predicting how diseases might jump from animals to humans, inspired by real-world crises such as COVID-19 and Ebola. However, turning that concept into a working machine learning system involved far more than I initially imagined.

One of the biggest lessons I learned was the importance of data ingenuity. The challenge of limited real-world data was initially daunting but by researching scientific papers, using proxy data and synthesizing scientifically plausible datasets, I developed a skill that goes beyond textbook machine learning: data storytelling. Learning how to simulate, validate, and justify synthetic data gave me more confidence as both a researcher and a data scientist.

I also deepened my understanding of model structuring and interpretability. Instead of simply building one predictive model, I created a system of models, each one representing a biological step in the outbreak process. This modular approach made debugging easier, improved transparency and resulted in stronger performance. It taught

me that sometimes the best solutions aren't the most complex, but they're the most aligned with the domain being modelled.

In the end, this project was more than a technical assignment. It was an opportunity to apply AI to a real-world issue, think critically, learn independently and build something that felt meaningful. It's given me both a deeper respect for interdisciplinary work and a clearer vision of the kind of problems I want to solve in the future, the kind that matter.

References

- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- Chen, Y., & Liaw, A. (2001). Random forests. R package version 4.6-14.
- Chua, K. B., et al. (2000). Nipah virus infection in bats (family Pteropodidae) in peninsular Malaysia. *Emerging Infectious Diseases, 6*(3), 309-312.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232.
- Han, B. A., et al. (2015). Zoonotic disease risks associated with wildlife and livestock trade in Southeast Asia. *The Lancet, 384*(9966), 1044-1049.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844.
- Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., & Gittleman, J. L. (2013). Global trends in emerging infectious diseases. *Nature, 451*(7181), 990-993.
- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods, 26*(6), 1481-1496.

- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22.
- Lin, S., et al. (2017). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 318-327.
- Luis, A. D., et al. (2013). A comparison of bat species as reservoirs for zoonotic viruses. *Emerging Infectious Diseases, 19*(2), 359-365.
- Meerburg, B. G., et al. (2009). Rodents as reservoirs of zoonotic diseases. *Vector-Borne and Zoonotic Diseases, 9*(2), 91-99.
- O'Shea, T. J., et al. (2014). Bats as reservoirs of emerging viruses. *Science, 346*(6211), 215-219.
- Olsen, B., et al. (2006). Global patterns of influenza A virus in wild birds. *Science, 312*(5772), 384-388.
- Patz, J. A., et al. (2005). Impact of regional climate change on human health. *Nature, 438*(7066), 310-317.
- Peuquet, D. J. (2002). *Representations of geographic space: A conceptual framework*. Springer.
- Smith, K. F., Goldberg, M. S., & Rosenthal, S. (2014). Predicting zoonotic spillover from wildlife and livestock. *Frontiers in Public Health, 2*, 1-13.