Flash Floods Prediction Model: A Case of Nairobi

Bico Steve
170504

Submitted in Partial fulfilment of the Requirements for the Degree of Master of Science in Computing and Information Systems at Strathmore University

School of Computing & Engineering Sciences
Strathmore University
Nairobi, Kenya

June, 2025

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of any degree by this or any other university. To the best of my knowledge, this dissertation contains no material previously published or written by another person, except for reference materials cited.

Student's Name: Bico Steve

Sign:  Date: 26th May, 2025.

Approval

This dissertation of Bico Steve was reviewed and approved by the following:

Prof. Ismail Ateya,
School of Computing and Engineering Sciences,
Strathmore University.

Dr. Julius Butime,
Dean, School of Computing and Engineering Sciences,
Strathmore University.

Prof. Bernard Shibwabo Kasamani,
Director of Graduate Studies,
Strathmore University.

Abstract

This study developed a flash flood prediction model for Nairobi, focusing on improving the city's resilience and emergency preparedness and response to flash floods. The study employed quantitative and experimental research design. A machine-learning model was built to predict flash flood occurrences using historical rainfall, soil moisture, and meteorological, hydrological, and topographical data. The key variables identified to influence flash flood occurrence were rainfall, soil moisture content, river discharge, and erosion degree, with rainfall showing the highest correlation (61%) to flash floods. Among various machine learning models, the Random Forest model outperformed others with an accuracy of 93.33%, recall of 90.47%, and an F1 score of 0.90, making it the most reliable predictor. Other models, such as KNN, Logistic Regression, SVM, and ANN, also showed impressive performance. The developed model can potentially improve flood prediction which can lead to reduction on damages, enhance Nairobi's resilience to flash floods, and providing a reference for other urban areas facing similar climate challenges. It was recommended that Nairobi and similar metropolitan areas should invest in enhancing their drainage infrastructure to complement the predictive model's capabilities. Integrating this model into city planning, emergency response systems, and early warning systems could help in mitigating the risks posed by flash floods. Additionally, policymakers should prioritize land use planning and environmental conservation to address the key drivers of flash floods, such as soil erosion and improper land use.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

3G	Third Generation
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
AUC	Area Under Curve
CPU	Central Processing Unit
DEM	Digital Elevation Model
EM – DAT	Emergency Events Database
GPU	Graphics Processing Unit
HEC-HMS	Hydrologic Engineering Centre's Hydrologic Modeling System
HEC-RAS	Hydrologic Engineering Centre's River Analysis System
LR	Long Range
LULC	Land User Land Cover
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
NBS	Nature-Based Solution
PCA-ANN	Principal Component Analysis and Artificial Neural Networks
RF	Random Forest
ROC	Receiver Operating Characteristic
SCADA	Supervisory Control and Data Acquisition
SR	Short Range
SVM	Support Vector Machine
SWMS	Storm Water Management Simulation

Definitions of Terms

Binary Encoding	Use of binary digits 0 and 1 as a fundamental unit of information for data points with a wide range (Todd, 2022).
Frequency Encoding	Categorical Representation through Counting (Ninja, 2024).
Hash Encoding	Process of converting data into fixed length string of letters and numbers (Team, 2023).
Hydrophobicity	Ability of soil to repel absorption of water ranging from seconds to weeks (Doerr et al., 2000).
Label Encoding	Assigning integers to different classes (Sonoda, 2024)
MaxAbsScaler	Data preprocessing method is used in machine learning to scale each feature by its maximum absolute value (Chouhbi, 2020a).
Mean Target Encoding	Method is used in machine learning to encode categorical variables (Kumar,2023).
MinMaxScaler	Machine learning data preprocessing method to scale features to a specified range, typically between 0 and 1 (Chouhbi, 2020b).
One-Hot Encoding	Process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions (Fawcett, 2024).
Ordinal Encoding	Method converts categorical data into numerical format by assigning a unique integer to each category (Brownlee, 2020a).
Receiver Operating Characteristic	Graphical Representation of performance of a binary classification model (Nahm, 2022).
RobustScaler	Data preprocessing technique used in machine learning to scale features in a robust way to outliers (Brownlee, 2020b).
StandardScaler	Data preprocessing method standardises features by removing the mean and scaling them to unit variance (Pal, 2023).

Acknowledgments

First and foremost, I thank the Almighty God for giving me the strength and opportunity to pursue this master's degree. I also wish to acknowledge and thank my family and friends for encouragement. Finally, I would like to appreciate Professor Ismail Ateya, for his guidance and assistance throughout this process.

Dedication

I dedicate this dissertation to my family, who have been there for me during the good and the bad times. Thank you for your patience and support throughout the journey.

Chapter 1: Introduction

1.1 Background of the Study

Flash floods are those more hazardous types of floods that can occur with little or no warning. Flash floods may result in water rising significantly in a short period (NOAA, n.d). In the recent past, due to climate change, there have been frequent flash floods with devastating consequences manifested in the degradation of infrastructures like roads, social amenities like parks, buildings, drainage, and affecting the population in general. A 2021 study by climate experts highlighted that climate change is worsening the effects of flooding, with human-caused warming leading to heavier downpours—by 3-19% in regions like Western Europe, where flash floods claimed nearly 200 lives (Shukla, 2021). By 2022, approximately 20 percent of Germany's population was at risk of flooding, while the Netherlands had the highest exposure in Europe, with around 60 percent of its population living in flood-prone areas.

In Asia, where many countries are prone to monsoon seasons, climate models predict more intense flooding and prolonged monsoons due to increased greenhouse gas emissions (Lai, 2023). South and Southeast Asia, with densely populated low-lying cities, are particularly vulnerable to these climate-related flooding risks, which are expected to intensify in the future. A study published in September 2021 by the World Weather Attribution says there is a direct link between climate change and the extreme rainfall events that led to deadly floods in the Mediterranean region, including Greece, Turkey, and Libya. Researchers determined that climate change has made these flooding events between ten and fifty times more likely and intense (Igini, 2023). The analysis, which involved examining historical rainfall data, climate models, and socioeconomic factors, revealed that increased greenhouse gas emissions have intensified the hydrological cycle, leading to more frequent and severe rainfall. The study highlighted that Greece experienced a 1-in-250-year flood event, made ten times more likely and 4 out of 10 more intense due to human-induced climate change. In Libya, typically experiencing floods once every 300 to 600 years, the likelihood increased up to fifty times, with intensity rising by 50% compared to a cooler climate.

In Eastern African countries, particularly Ethiopia, Somalia, and Tanzania, heavy rains, flash floods, and landslides caused significant disruption, with nearly 1.6 million people affected by May 30 2024 (OCHA, 2024). Among those affected, 528 people lost their lives, and approximately 482,320 individuals were displaced. While rains in Kenya, Somalia, and southeastern Ethiopia were projected to decrease in late May, the Climate Prediction and Applications Centre (ICPAC) announced on May 21, during the 67th Greater Horn of Africa

Climate Outlook Forum (GHACOF67), that other parts of the region were likely to experience above-average rainfall from June to September 2024 (Xinhua, 2024). Countries like Djibouti, Eritrea, central and northern Ethiopia, parts of Kenya, Uganda, South Sudan, and Sudan were forecasted to receive this heightened rainfall. This period was particularly critical in northern and western regions of the Greater Horn of Africa, as it often accounts for over 40 percent of annual rainfall and, in some areas, up to 90 percent. In Tanzania, heavy rains, landslides, and floods led to the loss of 155 lives and affected around 126,000 people by early May, according to local authorities (UNICEF, 2024). In Somalia, the Gu season rains (April-June) affected more than 268,000 people, displacing around 38,700 by mid-May 2024. Nine fatalities were recorded, and significant damage was done to schools, shelters, and infrastructure.

In Kenya, between March 1st and May 29th, 2024, 315 people lost their lives, 188 were injured, and 38 were reported missing due to heavy rains and flooding (Grignon, 2024). More than 306,520 people (61,304 families) were affected, including an estimated 293,200 people (58,641 families) displaced (Voice of America, 2024). Floods also caused significant property damage and disrupted essential services. Tropical Storm Ialy, which struck the coastline on May 21, brought heavy rains, strong winds, and high waves, resulting in two deaths and six injuries. Several health facilities across seven counties were either flooded or made inaccessible, impacting healthcare services. Nairobi has not been spared of these consequences with erratic weather patterns that are hard to predict which brings devastating effects to the residents of the city. Sometimes the city is too cold or too hot with excessive rainfall.

Apart from the climatic conditions, human factors like rapid population growth, increase in infrastructure development, and industrialization are among the factors that have led to social, environmental, and economic challenges. Urban development and industrialization have extensively changed the land use/land cover (LULC) and drainage pattern of the cities and urban centres (Zafar, 2024). According to Rukanga (2023), Nairobi with a population of over 4.5 million inhabitants, the sheer scale of the March/June 2024 rainfall has exposed the city's infrastructure and planning weaknesses. Another factor that has put Nairobi at a disadvantage is its location, which is on top of the Nairobi River flood plains, with other rivers flowing through the city. In addition, some of the rivers have dried due to construction and urbanization, but during flooding, the flood waters find the original course of the dried rivers. The infrastructure has also not been at par with the growth of the human population, with most drainages blocked or over-utilized.

Several studies have been conducted to handle flooding challenges globally by developing various forecasting models to handle flooding challenges globally. Lin et al. (2022),

in their research, developed a web-based prototype system for flood simulation and forecasting based on the HEC-HMS model. However, the study did not use parameters such as temperature, precipitation, and soil topography in its prediction. Rawas et al. (2024) reviewed innovative technologies such as artificial intelligence (AI)/machine learning (ML), the Internet of Things (IoT), cloud computing, and robotics used for flash flood early warnings and susceptibility predictions. This study primarily discussed flash flood technologies in general without a specific geographic focus, and it did not make use of climate change data in its prediction. The reviewed study mentioned the need to consider climate change impacts in future flash flood models.

To effectively address the shortcomings and research gaps identified in existing flood forecasting models, this study proposes the development of a web-based machine learning model that incorporates both climate data and soil topography. Unlike previous models that often relied solely on historical data and overlooked the dynamic impacts of climate change, this approach integrated real-time climate variable such as temperature, precipitation, and wind speed with critical soil characteristics, including erosion levels and sensitivity to capping. By harnessing advanced machine learning algorithms, the study aimed at enhancing the accuracy of flood predictions while providing adaptive insights tailored to the unique hydrological context of Nairobi. This innovative model will not only fill the existing research void by focusing on the interplay between soil and climate factors but also support decision-makers with timely, data-driven alerts for flash flood risks. Ultimately, this work strives to contribute to the development of resilient flood management strategies, ensuring communities are better prepared for future flash flood challenges.

1.2 Problem Statement

Flash floods are a frequent and destructive natural hazard that has resulted in many fatalities, property damage, disruption of daily lives, and loss of life nearly every year in Nairobi (Kilavi et al., 2018). The most affected areas are informal settlements because of unsupervised land use and the building of semi-housing structures (Kaburu et al., 2019). The 2024 floods significantly affected 43 counties, displacing 55,109 households and impacting a total of 101,132 households (Kenya Red Cross Society, 2024). The disaster led to 294 fatalities and 162 missing persons. Critical infrastructure was severely damaged, with 68 roads destroyed, 151 schools affected, and 1,373 businesses disrupted. Additionally, 11,539 livestock were lost, and 65,377 acres of crops were destroyed, posing a severe threat to food security.

Public health was compromised, with 45 health facilities affected and 2,458 water sources destroyed, further worsening the situation for the affected population.

Among the studies conducted in the line of flash floods, Kilavi et al. (2018) examined the correlation between multi-day heavy rainfall episodes and the active Madden–Julian Oscillation (MJO) phases alongside tropical cyclone activities, which contributed to enhanced moisture convergence over the region. This study utilized seasonal and short-term forecasting models to assess predictability. Wang et al. (2024) study innovatively combined deep learning methods, specifically a Temporal Convolutional Network (TCN), with flash flood simulation to predict the spatiotemporal dynamics of flash floods in complex terrains. The two studies are nearly identical to the current research in the sense that they are predicting flash floods. However, they used different prediction models and datasets to predict flash floods in different terrains from the current study.

The lack of a reliable and efficient localized prediction model that is adapted to Nairobi's geographic and climatic conditions is the problem this study seeks to address. While the Kenya Meteorological Department makes useful forecasts, weather predictions for localized phenomena such as flash floods prediction are inherently limited since the certainty of some of the warnings used are categorized as 'moderate possibility of occurrence' which makes the science of weather prediction imperfect (Mobanga & Amondi, 2024). The study aimed to develop a prediction model utilizing machine learning models and advanced data analytics for predicting flash floods that have higher precision. The proposed model aimed to bridge these gaps in prediction methods and to increase disaster preparedness, resulting in lower financial losses and protecting the Nairobians' lives and means of subsistence. The prediction model, therefore, could serve as a valuable basis for the Kenya Meteorological Department and the Kenya Red Cross Society in flood forecasting, disaster preparedness, and mitigation.

1.3 Research Objective

This study aimed to develop a machine learning model to predict flash floods in Nairobi.

1.3.1 Specific Objectives

i. To identify factors that influence the occurrence of flash floods in Nairobi County.
ii. To develop a machine learning model to predict flash floods in Nairobi County.
iii. To validate the accuracy of the developed model.

1.4 Research Questions

i. What are the factors that influence the occurrence of flash floods in Nairobi County?
ii. How can a machine learning model be developed to predict flash floods accurately?
iii. How effective is the developed model to predict flash floods?

1.5 Justification

Rapid urbanization in the city has seen a decrease in the quality of infrastructure, such as natural drainage and waste management systems. With the increased population, the city's infrastructure is overloaded and is even worse during rainy seasons. These heavy rainfalls have caused frequent flash floods that interrupt the daily activities of the city dwellers, destroy property, and put the residence health at risk from water borne diseases and other climate-associated diseases. A flash flood predicting model tailored towards Nairobi can provide timely warnings, allowing for better preparedness and response time and mitigating the adverse impact of flash floods.

The beneficiaries of the flash flood prediction system in Nairobi are multifaceted, encompassing both local authorities and the general population. Government agencies, such as Nairobi County's disaster management and environmental departments, will benefit significantly from the predictive capabilities of this model. With the system providing early warnings based on real-time data, these agencies can proactively allocate resources, coordinate emergency response teams, and issue timely public alerts. This not only mitigates the damage caused by flash floods but also reduces the risk of casualties, infrastructure damage, and long-term economic impacts. Furthermore, the data-driven insights from this model will enhance urban planning efforts, allowing authorities to identify high-risk areas and implement flood-prevention measures in future development projects.

The Red Cross, notably the Kenya Red Cross Society, will also be a key beneficiary of the flash flood prediction system in Nairobi. As a humanitarian organization dedicated to disaster response, the Red Cross Society plays a critical role in providing emergency relief, medical care, and shelter to those affected by floods. By integrating the predictive model into its operations, the Red Cross Society can significantly enhance its preparedness and response capabilities. Early warnings will allow the organization to mobilize volunteers, pre-position essential supplies, and coordinate with local authorities to reach vulnerable populations more efficiently. This proactive approach will not only save lives but also ensure that emergency interventions are timely and effective, minimizing the chaos and delays typically associated with sudden flash floods. Moreover, the Red Cross Society can use the system to coordinate

evacuation efforts and shelter management better. In areas at high risk of flooding, the organization can issue evacuation alerts to residents ahead of time, directing them to safer locations. Having accurate flood risk data allows the Red Cross to set up and manage temporary shelters in safer zones, ensuring that displaced families have access to adequate shelter, food, water, and medical care. This advanced planning, enabled by the flood prediction system, reduces the strain on emergency services and improves the overall effectiveness of disaster relief efforts.

Additionally, the scientific community gets to benefit from this study as the factors identified as causes of flash floods and the model created will add to the body of knowledge by integrating data sources like meteorological data, topographical information, and flood historical records. The model will use machine learning to increase the precision and dependability of flash flood prediction.

1.6 Scope and Limitations

The scope of this study focused on developing a flash flood prediction model for Nairobi, using a combination of meteorological, hydrological, topographical, and geospatial data to identify key drivers of flash floods. The study aimed to improve the city's resilience and emergency response to such events through machine learning techniques, with a primary focus on predicting flash floods in urban areas.

The study has several limitations. First, the model's performance is dependent on the availability and accuracy of historical data, which may be limited in some regions. Additionally, the study focused on flash floods in Nairobi and may not fully account for the unique challenges posed by different urban environments. The model's effectiveness may vary in other cities with different topographies, climates, and infrastructure. Finally, while the study highlighted the importance of certain variables, it does not address all possible contributing factors to flash floodings, such as human behaviour or long-term climate change impacts.

Chapter 2: Literature Review

2.1 Introduction

Flash floods are sudden and severe floods that can devastate infrastructure, disrupt daily life, and endanger human safety. Rapid urbanization, inadequate drainage systems, type of soil, and climate change all exacerbate the impact of flash floods in cities such as Nairobi, Kenya's capital. The increasing frequency and intensity of these events necessitate the creation of reliable prediction models to mitigate their negative consequences. This literature review will look at existing studies on flash flood prediction models, with a focus on how their application can be relevant in Nairobi. Various methodologies, technologies, and case studies will be reviewed to identify the most effective approaches for predicting flash floods in an urban African context. Challenges and limitations of these models will also be discussed, as well as potential areas for future research. The review will also address the models' challenges and constraints.

2.2 Empirical Literature

In the recent past and currently, climate change has manifested everywhere in the world, with melting glaciers in the Arctic, extreme heat waves in Europe, the Middle East, and the Americas, as well as heavy erratic rainfall in Africa and the Caribbean. While climate change can occur naturally, the current rapid trends have been caused by human activities (BBC News,2024). Among the effects of climate change are flash floods, and with this problem, many studies and experiments have been carried out to make predictions on flash floods and flooding in general in various cities around the world. The most notable studies have been done in Chinese cities, which have experienced heavier-than-usual floods that have led to the destruction of properties and loss of lives.

This is a good initiative and leads to other cities or urban areas facing the same problem. The issue is that, sometimes, many of the countries and towns affected by flash floods are not aware of the causes or how to mitigate and minimize the flash flood damage. In the causes where they have information regarding how to create some sense of mitigation, they lack the technical know-how or even funds, rendering their methods ineffective and hence no consequence in the idea of prevention or proper mitigation. The effective of flash floods on infrastructure and the population in general cannot be underestimated. To confront this issue, there is a need to create solutions that are workable and customizable to the specific region that is facing the challenge.

Wu, Bhattacharya, Xie, and Zevenbergen (2023) proposed an improved flash flood forecasting method by integrating Flash Flood Guidance (FFG) with a Frequentist approach. The study analysed rainfall data from the Posina River basin in Italy, focusing on 94 six-hourly rainfall events, including 23 flood events. The methodology involved calculating deviations from log-transformed rainfall data leading to flash floods and using Kernel Density Estimation (KDE) to model these deviations. The output was fitted to a Normal Distribution Function (NDF) to calculate flood probabilities. The study introduced three probability thresholds (10%, 20%, and 60%) to classify flood risks into four categories: very low, low, significant, and high, with corresponding colour codes for straightforward interpretation. The findings showed that the Frequentist FFG method provided a range of flood probabilities (0% to 100%), improving accuracy in predicting flood risks and reducing false alarms compared to the traditional binary FFG system. This probabilistic approach aided decision-making by offering a more precise assessment of flash flood risks.

A study on improving flood forecasting developed a web-based prototype system (WSFF) based on the Hydrologic Engineering Centre Hydrologic Modeling System (HEC-HMS) model to address simulation challenges in web environments. The study used a dataset from the Chuanchang watershed in southeastern Fujian Province, China, to evaluate the WSFF's performance. A hydrological model of the Chuanchang watershed was built, and empirical equations incorporating key parameters λ and V were established using data from nine historical flood events. These models and parameter identification knowledge were then integrated into the WSFF. The system's accuracy was tested with three additional flood events. The WSFF demonstrated consistent and satisfactory performance in forecasting peak flow, total flood volume, and peak flow timing. The average Nash-Sutcliffe efficiency (NSE) was 0.81 for validation events and 0.82 for calibration events. The relative error in peak flow (REP) was within 15%, and the timing error (RET) was within 1 hour, indicating reasonable accuracy. The WSFF, developed as open-source software, serves as a valuable tool for the hydrological community in applying HEC-HMS to flood forecasting efforts (Lin et al., 2022).

A study by Al-Rawas, Nikoo, Al-Wardy, and Etri (2024) reviewed the application of emerging technologies for flash flood prediction and early warnings, including artificial intelligence (AI), machine learning (ML), the Internet of Things (IoT), cloud computing, and robotics. The study analysed articles published between 2010 and 2023 from databases like Google Scholar, Scopus, and Web of Science. The findings indicated that AI/ML technologies were the most used, applied in 64% of the studies, followed by IoT (19%), cloud computing (6%), and robotics (2%). Standard AI/ML methods included random forests and support vector

machines with high prediction accuracy (ROC and AUC > 0.90). However, these models require further optimization and larger test datasets for improvement. The integration of AI/ML, IoT, and cloud computing enables real-time dissemination of early warnings through platforms like SMS and social media, although issues with internet connectivity and data loss persist. AI/ML models commonly use topographical, geological, and hydrological variables for susceptibility prediction, but the variable selection lacks theoretical consistency. Future studies should consider incorporating sociodemographic, health, and housing data for more reliable flood risk assessments, as well as projections under different climate change scenarios to aid long-term adaptation strategies.

According to the study conducted by Wang et al. (2024), a novel approach combining deep learning methods with hydrodynamic models was developed to improve flash flood simulation, particularly addressing challenges such as data scarcity, small sample sizes, and complex terrain. The study proposed a Temporal Convolutional Network (TCN) model to predict the spatiotemporal dynamics of flash floods. Typical rainfall patterns were extracted from the study area, and a hydrograph dataset was generated using design storm methods, incorporating various rainfall patterns and return periods. The TCN model's performance was benchmarked against a Convolutional Neural Network (CNN). The findings revealed that the TCN model effectively predicted flash floods, with average Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Nash-Sutcliffe Efficiency (NSE) values of 0.04, 0.17 and 0.834, respectively, on the validation set. Although the CNN model performed better in small flood scenarios, the TCN model showed better stability, fewer outliers, and higher consistency during the flood's recession period. The TCN-based rapid simulation method demonstrated strong capability in capturing the dynamic characteristics of flash floods, particularly in mountainous areas, offering a promising new method for flash flood prediction and early warning.

In a study by Kilavi et al. (2018), the exceptionally wet Long-Rains season (March-May) of 2018 in Kenya was examined to understand the causes, impacts, and predictability of extreme rainfall events, with a focus on implications for flood risk management. The study found that the unusually high rainfall totals in March and April resulted from multiple multi-day rainfall episodes rather than single extreme daily events. Three significant intra-seasonal rainfall events led to widespread flooding, causing loss of lives, displacement, infrastructure damage, and disruption of essential services. These events were linked to active phases of the Madden–Julian Oscillation (MJO) and tropical cyclones over the southwest Indian Ocean, which drove moisture convergence and enhanced convection over Kenya. Predictability of the

rainfall events was assessed, revealing that while long-lead seasonal forecasts showed limited ability to predict heavy rainfall, sub-seasonal and short-term forecasts provided more precise signals of extreme weather, mainly due to skill in predicting MJO activity. The study highlights the need for integrating sub-seasonal and short-term forecasts to improve flood preparedness in Kenya, as well as the importance of differentiating between forecast lead times for Kenya's two wet seasons. Enhanced early warning and action systems for flood management were recommended, especially given the short anticipatory window during the long rain season.

In a study conducted by Tom et al. (2022), urban flood modeling in developing cities, with a focus on informal settlements such as Mukuru slums in Nairobi, Kenya, was reviewed. Informal settlements, characterized by unplanned housing and inadequate infrastructure, face heightened flood risks that have been underexplored compared to rural flooding. The study emphasizes the need for flood modeling as a key component of integrated flood risk management in urban settings. A desk review was conducted to explore various flood models, their strengths, limitations, and the role of model calibration and stacking in addressing uncertainties and capturing flood dynamics at multiple scales. Model stacking was highlighted to understand flood risks across different city scales. The review gathered literature from sources such as Google Scholar, ScienceDirect, and Elsevier, focusing on studies from 1990 onwards. The study underscores the growing flood risks in informal settlements as cities expand and call for better flood modeling practices to manage these risks effectively. Figure 2.1 shows events that will lead to flooding in urban areas.

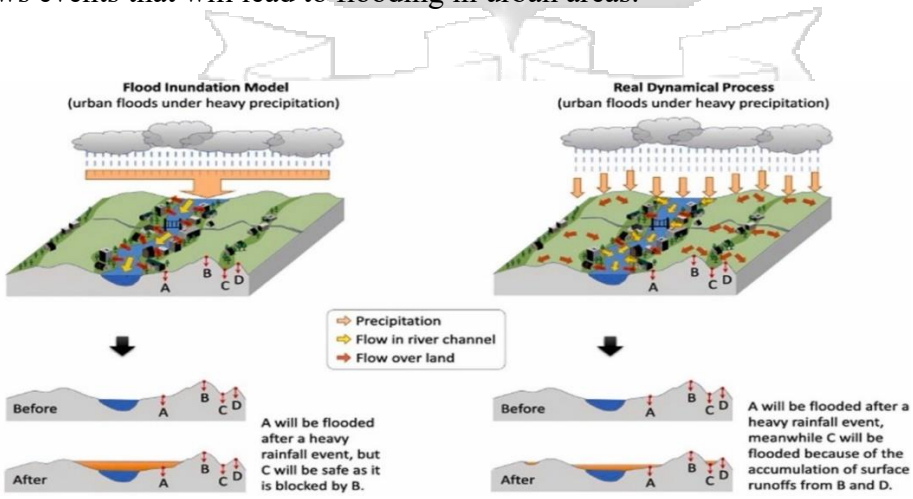


Figure 2.1: Schematic representation of flooding in urban areas (Tom et al., 2022)

2.3 Theoretical Literature

2.3.1 Types of Floods

There are two primary types of floods: flash floods and river floods. A flash flood is a flood that begins within 6 hours, and often within 3 hours, of a heavy rainfall (NOAA, n.d). Flash floods occur when water levels rise quickly due to heavy rain, especially in low-lying areas. These events are hazardous because of their sudden onset and the speed at which they can develop, often leaving little time for people to seek higher ground or take protective actions. Flash floods are widespread in regions with arid climates and rocky terrain, where the absence of soil or vegetation reduces the natural barriers that could slow or absorb the flow of rainwater. Flash floods can result from a variety of causes, but they most commonly occur due to intense rainfall from thunderstorms. In some cases, they can also be triggered by dam or levee failures, as well as mudslides or debris flows. Several factors influence how quickly flash flooding may develop and where it occurs, including the intensity and location of rainfall, land use, topography, vegetation types and density, soil type, and the soil's water content (National Severe Storms Laboratory, n.d). Urban areas are especially susceptible to rapid flooding. Rainfall from the same storm can lead to more severe flooding in cities compared to suburban or rural areas. This is because impervious surfaces, such as roads and buildings, prevent water from soaking into the ground, causing it to flow into low-lying areas quickly. Flash floods happen so swiftly that they often catch people off guard. If individuals are travelling, they can face dangerous conditions if they encounter high, fast-moving water. At home or in businesses, the rising waters can trap people or cause property damage before they have a chance to respond or take protective measures.

On the other hand, river flooding happens when the water in a river exceeds the capacity of its banks, causing it to overflow. This type of flooding is more frequent in regions with wetter climates, extended rainy seasons, or areas near melting snow and ice, where the accumulated water cannot be held within the river's channel (Zurich Insurance Group, 2024). River floods can cause widespread destruction as the overflow impacts smaller rivers downstream, potentially leading to dam and dike failures that flood surrounding areas. To assess the likelihood of river flooding, models consider historical and forecasted rainfall, current river levels, and soil and terrain conditions. The severity of a river flood depends on the landscape and the duration and intensity of rainfall in the river's catchment. Additional factors include soil saturation and the effects of climate change on rainfall patterns. In flat areas, floodwaters tend to rise slowly and remain shallow but may persist for several days. Conversely, in hilly or

mountainous regions, floods can occur quickly after heavy rainfall, drain rapidly, and cause significant damage due to debris flow.

2.3.1 Causes of floods

Several studies have been conducted to determine the causes of flooding in urban areas and cities. Among the causes mentioned are inadequate planning, lack of proper provisions of services and equipment, focusing on areas that are vulnerable to flooding, and extreme increases in rainfall frequency that have been attributed to climate change (Ramiaramanana & Teller, 2021). According to Ferreira et al. (2010), growth in urbanization poses several risks related to water management issues, such as food shortages and deterioration of water quality, which renders the population vulnerable. Because of the harsh weather and hydrometeorological circumstances, Ferreira et al. (2021) state that the increase in urbanisation possess several risks among the water management problem like degradation of water quality, therefore making the population vulnerable. Urbanization and global warming exacerbate flooding because of high and extreme hydrometeorological conditions. Today, 56% of the world's population lives in urban centres and the percentage is projected to increase to nearly 68% by 2050 driven by urbanization trends especially in Africa and Asia (United Nations, 2018; World Bank, 2023). This population pressure, coupled with inadequate facilities in third-world cities like Nairobi, has made it very hard to control flash floods in the cities because the abandoned wetlands or rivers beds are used by those who do not get job opportunities to build informal housing.

Other scholars argue that floods are caused by drainage basin conditions, soil characteristics and permeability, soil moisture content and its vertical distribution, rate of urbanization, and the presence of dykes, dams and reservoirs (Kundzewicz et al., 2013). Brackenridge et al. (2012) argue that floods are caused by extreme storm surges or extreme tide events, but this can only apply to areas around bigger lakes or oceans. It seems from these scholars that they agree on the causes of flooding, with significant note being rapid urbanization.

2.3.2 Effects of the floods

Flooding, particularly flash floods, can wreak havoc on entire cities and urban landscapes. Historically, rapid flooding has claimed many lives, often resulting in secondary disasters such as landslides and infrastructure collapses. One of the most severe consequences of flooding is the loss of homes and personal property, along with the incapacitation of vital

buildings and infrastructure, including hospitals and care facilities for the elderly. Disruptions to power and mobile communication are common during floods, which can hinder access to safety and affect people's livelihoods.

The economic impact of floods can be substantial, significantly affecting critical industries and sectors such as agriculture, fishing, food production, healthcare, labour, and tourism. Research suggests that recurrent flooding could reduce a region's GDP by as much as 11% by the end of the century (Lai, 2023). Recovery from such resource losses can take countries years. Individuals residing near rivers, in wetter climates, or areas prone to monsoon seasons are particularly vulnerable to flooding. Countries in South and Southeast Asia, such as Bangladesh which experienced a third of its land submerged at one point in 2020 and India, have faced severe flood events in recent years due to their low-lying geography and dense populations (Mishra, 2024). As a result, there has been significant mass migration and population displacement over the past few decades, leading to overcrowding in urban areas and an increase in urban poverty. This trend raises concerns about long-term social inequalities and potential unrest.

Flooding has been a recurring problem in Nairobi, Kenya, wreaking havoc on the city's infrastructure, economy, and citizens' daily lives. Nairobi has seen regular flooding, with significant floods happening about every two years on average. Between 1964 and 2004, Kenya saw 17 catastrophic floods, each impacting around 70,000 people. Climate change, urbanization, and insufficient drainage systems have all had an impact on how frequently these disasters occur (Kilavi et al., 2018). Significant loss of life and displacement have occurred due to floods. For example, in recent years, heavy rains and floods have killed hundreds of people and displaced thousands. In May 2024 alone, 267 people died, and 281,835 people were displaced because of persistent heavy rains (OCHA, 2024). In addition, there has been significant damage to infrastructure, such as roads, bridges, and buildings. The consequences of flooding on the economy are substantial. Floods cause livestock losses, crop destruction, and disruptions to commercial operations. Nearly 10,000 animals perished, and 41,562 acres of crops were damaged in 2024. The financial resources of the city are further taxed by the expense of reconstructing and fixing damaged infrastructure.

Rukanga (2024) states that health and sanitation problems are made worse by flooding, especially in informal settlements. Contaminated water sources and overflowing sewers increase waterborne disease risk. Particularly vulnerable are slum areas like Mukuru and Mathare due to inadequate waste disposal and drainage systems—environmental repercussions, including biodiversity loss and soil erosion. The natural absorption capacity of

the land has been diminished by urban development's encroachment on riverbanks and wetlands, making the flooding situation worse.

2.3.3 Triggers to surface water runoff

Soil water repellence (hydrophobicity) is the reduction of the affinity of soil to retain water such that the soil resists water wetting for long periods ranging from seconds to even weeks. This, in effect, reduces the infiltration capacity of the soil, which enhances flow and accelerates soil erosion, uneven wetting patterns, and leeching of agrochemicals. They go ahead and state that clay and sand percentage determine the infiltration and water-holding capacity of the soil. High clay content can be related to high water retention capacity and vice versa. Aggregated clay content causes soil to be fractured, which promotes drainage under wet conditions. In addition, canopy and tree covers also affect soil moisture content, with high vegetation cover reducing the surface runoff but negating evaporation due to shading. This, in contrast, is reversed by transpiration from the vegetation leaves (Doerr et al., 2000).

Cheng et al. (2016) maintain that Nairobi areas have block cotton soil, which shrinks or swells depending on weather conditions. In his text, Mwangangi (2023) asserts that many places in Nairobi and the satellite towns around it have blocked cotton soil. He goes on to say that black cotton soil has high plasticity because of the high clay content, which makes it retain and absorb a high amount of water and retain it for an extended period, especially during wet seasons. Because of the high retention of water, the soil has low water permeability, which causes high water stagnation. The black cotton soil in Nairobi and its environment makes the region experience flash floods.

2.3.4 Other Solutions

Among the notable mentions of alternative ways of predicting floods is the Nature-Based Solution (NBS), which was suggested by Ferreira et al. (2021). They assert that there is a need to use vegetation to control flooding by planting trees and grass in flooding zones. The main problem with this suggestion is that it is influenced by design and placement aspects. Berlin and other Chinese cities have implemented these solutions. Examples of this include reforestation and afforestation, wetland restoration, green infrastructure, and sustainable land management (Pakistan Department of Economic and Social Affairs, 2023). Bibi et al. (2023) suggest Storm Water Management Simulation (SWMS). They claim the drainage system malfunctions due to the impervious nature of cities' soils and variations in rainfall intensity. In their argument, these methods check land use and climate change impact. From this test, peak

surface run off, infiltration rate, and flooding volume can be measured. The drawback with this solution is that it is more of a reactive approach than a proactive approach, and by the time the solution comes to those affected, the damage will have already been done.

There is another method of predicting flash floods in cities. It involves measuring the varying water levels in rivers in the towns and streams. The process uses Supervisory Control and Data Acquisition (SCADA) techniques. Here, computers with sensors are used with the sensors placed in the flowing riverbeds. The computers are networked with 3G, which means when there is a flood, the river level swells, and the colour of the water turns muddy. This, in return, notifies the populace of the danger coming (Achakakorn et al., 2014; Intharasombat & Khoenkaw, 2015). This method has a weakness in that it can only be used in situations where there is a river flowing within a city. In addition, there might be inaccuracy in river colour water, or the sensor might send false alarms due to winds.

Pannamperuma and Rajapakse (2020) suggest the use of the Digital Elevation Model (DEM). In this method, GIS software is used to model ponds to control floods. Excessive rainwaters are directed to ponds, which are strategically located to avoid overflow in pavements and residential areas. The weakness of this method is that it is costly to implement and not ideal for a city like Nairobi, which is a low-lying water catchment zone. Among the many solutions mentioned is the Synthetic Aperture Radar method. This involves the processing of images of a region from two separate dates and then comparing to check whether water and non-water differentiation. From the images of the result, predictions will be made as to whether there is flooding or the absence of floods. The disadvantage of using this method is that it can be affected by localized factors like tall buildings' shadows and vegetation, which may make the results inaccurate. It also ignores hydrometeorological conditions and focuses primarily on human factors (Jardosh et al., 2020).

2.3.5 Machine Learning Models

When A.L. Samuel, an American computer scientist, developed a self-learning checkers software in the late 1950s, he came up with the phrase ‘machine learning.’ However, it was not until the 2000s that machine learning became a widely accepted term thanks to improvements in processing power and greater accessibility of massive data. According to Mitchell (1997), machine learning is a subset of artificial intelligence. Machine learning, according to Ratner (2017), is the ability of a machine to learn a data structure without explicit programming. As a result, machine learning uses statistical models and algorithms to perform various mathematical operations on collected data and discover patterns and relating within

them. In conclusion, sampling data, also referred to as training data, is the primary goal of machine learning. Consequently, statistical theory is used in machine learning to create models that depict patterns in data that, if found, can generalize complicated problems.

Machine learning (ML) and artificial intelligence (AI) have emerged as effective tools for forecasting flash floods. Random forests, support vector machines, and neural networks have been used to predict flash flood risk and provide early warnings. These models can analyse large datasets and detect complex patterns that traditional models may overlook. AI/ML models have been shown in studies to be highly accurate, with receiver operating characteristics (ROC) and areas under the curve (AUC) greater than 0.90. However, these models require a large amount of training data as well as computational resources (Al-Rawas et al., 2024).

According to Alpaydin (2014), the model may be applied to description, prediction, or both. Descriptive models aid in describing and understanding a situation by illuminating the structure of the facts, whereas predictive models forecast future conditions based on data that characterize previous events (Witten, Frank, Hall, & Pal, 2017). The two primary categories of machine learning are distinguished by the various data input architectures that are utilized. These consist of both supervised and unsupervised learning. Supervised learning is a machine learning paradigm in which the model is trained using a labelled dataset. This implies that each training example is associated with an output label. The model's goal is to learn the mapping from inputs to outputs so that it can predict the outcome for new, previously unseen data. Supervised learning is commonly used for tasks such as classification and regression, whereas unsupervised learning involves training a model on data with unlabelled responses. The goal is to find hidden patterns or intrinsic structures in the input data. This type of learning is commonly used for clustering, association, and dimensionality reduction.

2.3.5.1 Decision Trees

Decision trees are a supervised learning algorithm that can be used for classification as well as regression. They split the data into subsets based on the value of input features, resulting in a decision tree-like model. Each internal node represents a "test" of an attribute, each branch represents the result of the test, and each leaf node represents a class label or a continuous value. The concept of decision trees originated in the 1960s, with early work by Hunt, Marin, and Stone. Breiman introduced the most notable algorithm, CART (Classification and Regression Trees), in 19841. Since then, decision trees have evolved dramatically, with numerous variations and improvements (Blockeel et al., 2023). Decision tree analysis entails visually depicting the potential outcomes, costs, and consequences of a complex decision.

These trees are beneficial for analysing quantitative data and making numbers-based decisions. Figures 2.2 and 2.3 show how decision trees work and how they can be used in flood prediction.

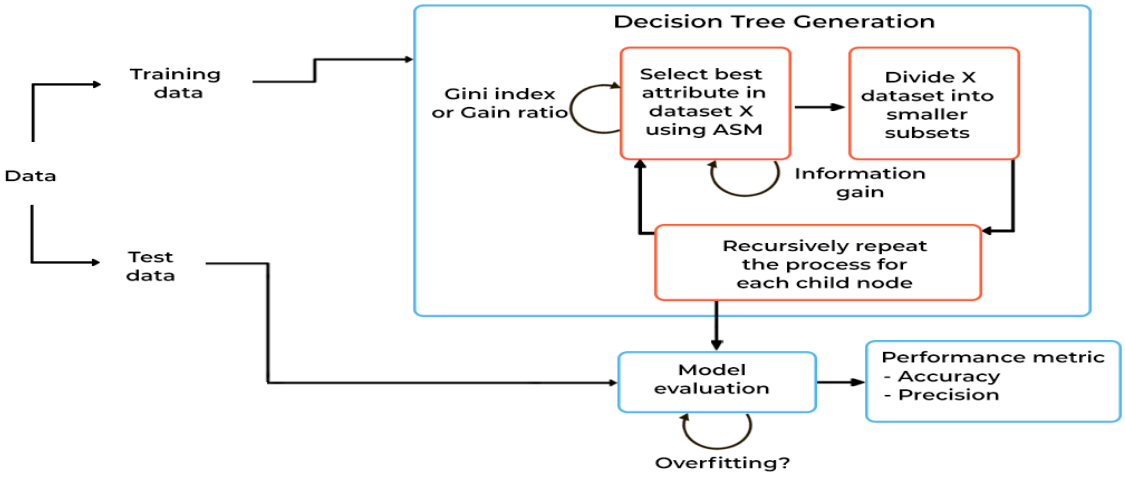


Figure 2.2: How Decision Tree works (Kanade, 2022)

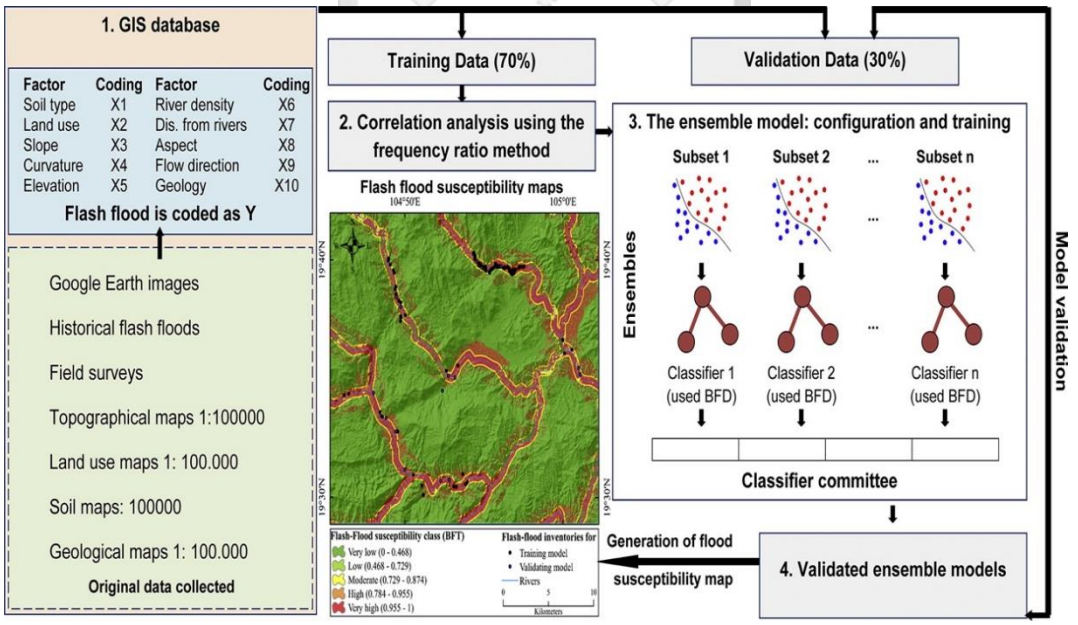


Figure 2.3: Use of decision tree in predicting floods (Pham et al., 2021)

2.3.5.2 K-Nearest Neighbours

The k-nearest neighbors (k-NN) algorithm is a fundamental machine learning method that is well-known for its simplicity and effectiveness in a wide range of applications. The k-nearest neighbors' algorithm is a nonparametric method for classification and regression. It works by identifying the k closest training examples in the feature space to a given test instance and making predictions based on their proximity. K-NN's simplicity and intuitiveness make it a popular choice for a wide range of practical applications. This algorithm was used in Bangladesh by Gauhar et al. (2021) to make flood predictions for the Bengal River. The

principles use the Pearson correlation coefficient (r_{xy}), which measures the strength of existing linear relations between two variables (e.g., x and y).

$$r_{xy} = \frac{\Sigma_{i1}^{n}(x_{i^-}\bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i1}^{n}(x^-_{i^-X)^2\Sigma_{i1}^{n}(y_i-\bar{y})^2}}}$$

Where n, x_i, y_t, and y respectively represent several data points, value of x (i^{th} observation), with the mean of x, value of you (for i^{th} observation), and mean of y. Spearman's rank correlation coefficient (r_s) is a measure of strength and direction between any of the two ranked variables.

$$r_s = 1 - \frac{6\Sigma_{i1}^{n} d_i^2}{n(n^2-1)}$$

Where n and d_i denote the number of cases and the difference between paired ranks, respectively, after correlation analysis, the following equation is used for feature selection.

$$AND(r_{xy}, r_s) \geq 0.50$$

2.3.5.3 Hybrid Models

Combining multiple techniques, processes, or algorithms to maximize their benefits and minimize their downsides is the goal of hybrid machine learning models. Performance, accuracy, and resilience may all be enhanced by this approach (Ankita Bhattacharya, 2022). One of the hybrid models that stands out is architectural integration, which mixes two or more classical algorithms. The Adaptive Neuro-Fuzzy Inference System (ANFIS), which combines artificial neural networks and fuzzy logic, is an example. An additional approach is the information control model, which combines conventional machine learning techniques with data manipulation techniques. One example of this type of model is the PCA-ANN Hybrid Model, in which an Artificial Neural Network (ANN) is used after Principal Component Analysis (PCA) reduces the dimensionality of the data. The case Adaptive Neuro-Fuzzy

Inference System was used by Nhita et al. (2015) for rainfall forecasting in Bandung. Figure 2.4 shows the flow of the ANFIS.

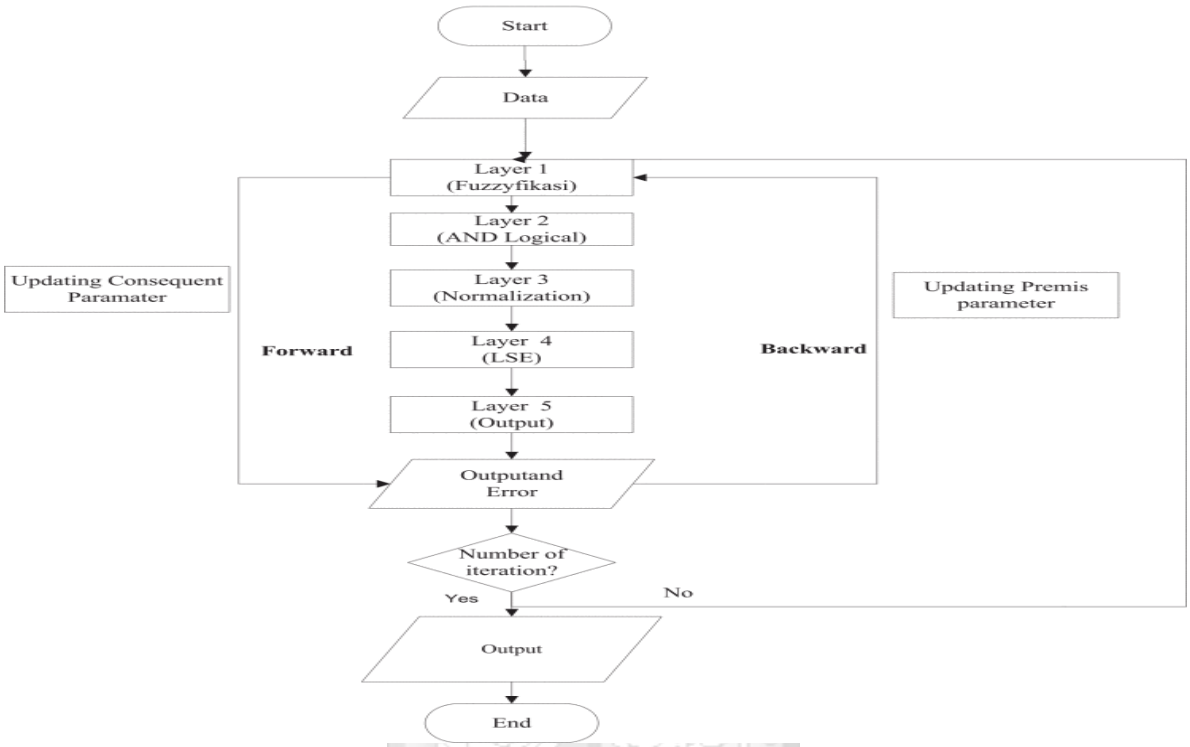


Figure 2.4: ANFIS (Nhita et al., 2015).

2.4 Frameworks

With the use of a machine learning framework, data scientists and software engineers may build machine learning models without having to figure out the underlying statistical and mathematical foundations of the algorithms. It makes development easier by removing the need for programmers to start from scratch every time they create a specific application. Machine learning frameworks comprise several related working libraries that facilitate the development of machine learning models, as will be covered in more details in the following sections.

2.4.1 PyTorch

PyTorch is a Python programming library designed to support deep learning applications. Deep Learning for Computer Vision, an expert approach to training complex neural networks using TensorFlow and Keras, provides responsiveness and ease. PyTorch facilitates deep neural network development. It is used for many different purposes and has a wide range of applications. PyTorch, like Python for programming, is a tool that can be used in expert real-world applications and serves as an excellent introduction to deep learning (Imambi et al., 2021). Figure 2.5 shows the PyTorch Model.

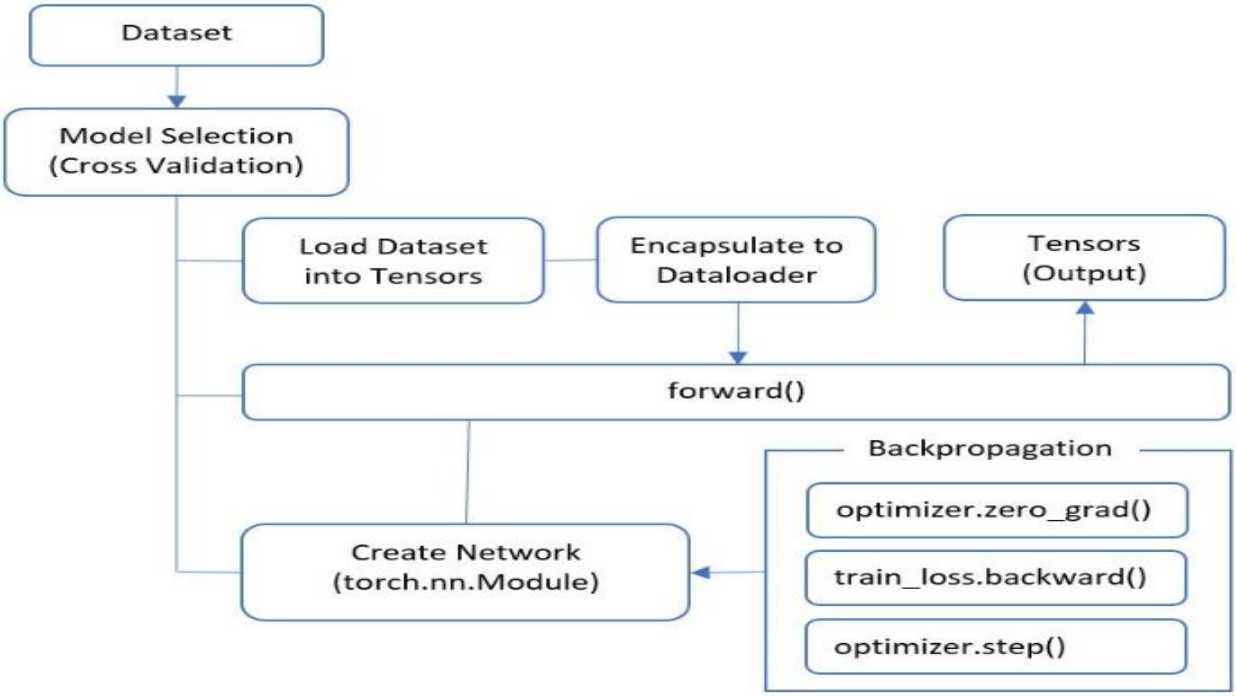


Figure 2.5: The Flow Diagram of the PyTorch Model (Hoc et al., 2023).

2.4.2 Keras

Bindings to deep learning libraries such as TensorFlow, CNTK, Theano, and the recently released Deeplearning4j are made possible via the Python wrapper framework Keras. It is available for free under the MIT license and was created for quick testing. Keras can function equally effectively on central processing units (CPUs) and graphics processing units (GPUs) because of their underlying frameworks (Nguyen et al., 2019).

2.4.3 Scikit Learn

The most reliable and practical Python machine-learning library is called Scikit-learn or Sklearn. A Python consistency interface offers a range of effective tools for statistical modeling and machine learning, including regression, clustering, classification, and dimensionality reduction. This library is based on NumPy, SciPy, and Matplotlib and is mostly developed in Python (SciKit Learn Tutorial, n.d.). According to Abraham et al. (2014), a two-dimensional array of size-sampled features can be used as data for any object or technique using scikit-learn. Its wide application across all disciplines is provided by this convention. All Scikit-learn objects share this consistent set of operations: Estimators can apply models to data, Predictors can create predictions based on fresh data, and Transformers may convert data between different representations.

2.5 Gaps in the Existing Systems

Even with the latest technological developments and technical expertise in weather pattern prediction, there are still gaps in the models used to predict flash floods, particularly in Africa where there is a dearth of data and literature about system implementation. To begin with, the River Nzoi Basin is the only significant prediction model in Kenya that has been reported, leaving other regions without any predictions. Additionally, Nairobi's climate is not comparable to that of the Nzoia basin. Secondly, the advanced models, particularly those developed by the Chinese, are not adaptable to Nairobi; as a result, they are unable to accurately forecast flash floods in Nairobi. Implementing some of the recommended strategies will cost money. For instance, they need costly devices, gear, and networking capabilities, all of which are difficult to maintain and buy in Nairobi, a Third World city. Furthermore, the security of these devices cannot be assured. Lastly, some researchers have focused on just one aspect of weather measurement, such as rainfall, failing to consider the multitude of variables that might cause surface runoff. This research did not consider several aspects, including plant cover, soil moisture content retention patterns, slope patterns, human activities, and soil type. In summary, a user-friendly, safe, and customizable paradigm is required for usage in cities, particularly in developing nations like Kenya.

2.6 Conceptual Model

The problem at hand is rationalized in the conceptual framework, along with a suggested solution that the study intends to implement. Through the user interface, the required dataset will be entered by the user. After querying the trained models for flash flood prediction, the user will be provided with a decision from the three best performing models. K Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVC), Artificial Neural Network (ANN), and Linear Regression (LR) algorithms will be used to train and assemble the model. Slope position, surface stoniness, area affected, erosion degree, sensitivity to capping, wind speed, temperature, soil moisture percentage, humidity, rainfall, river discharge, land use type, and elevation features will be used to train the models, which will then be combined to link across columns with various surface shapes. An assessment of the model's prediction will be given showing the categories of flash flood event. All these applications are required to ensure a seamless pipeline that ingests raw data, processes and splits it into training and testing sets, applies various machine learning models through a user-friendly prediction interface, and delivers accurate and interpretable prediction results to the user by integrating data handling, model inference, and database interaction. Figure 2.6 depicts the model's solution.

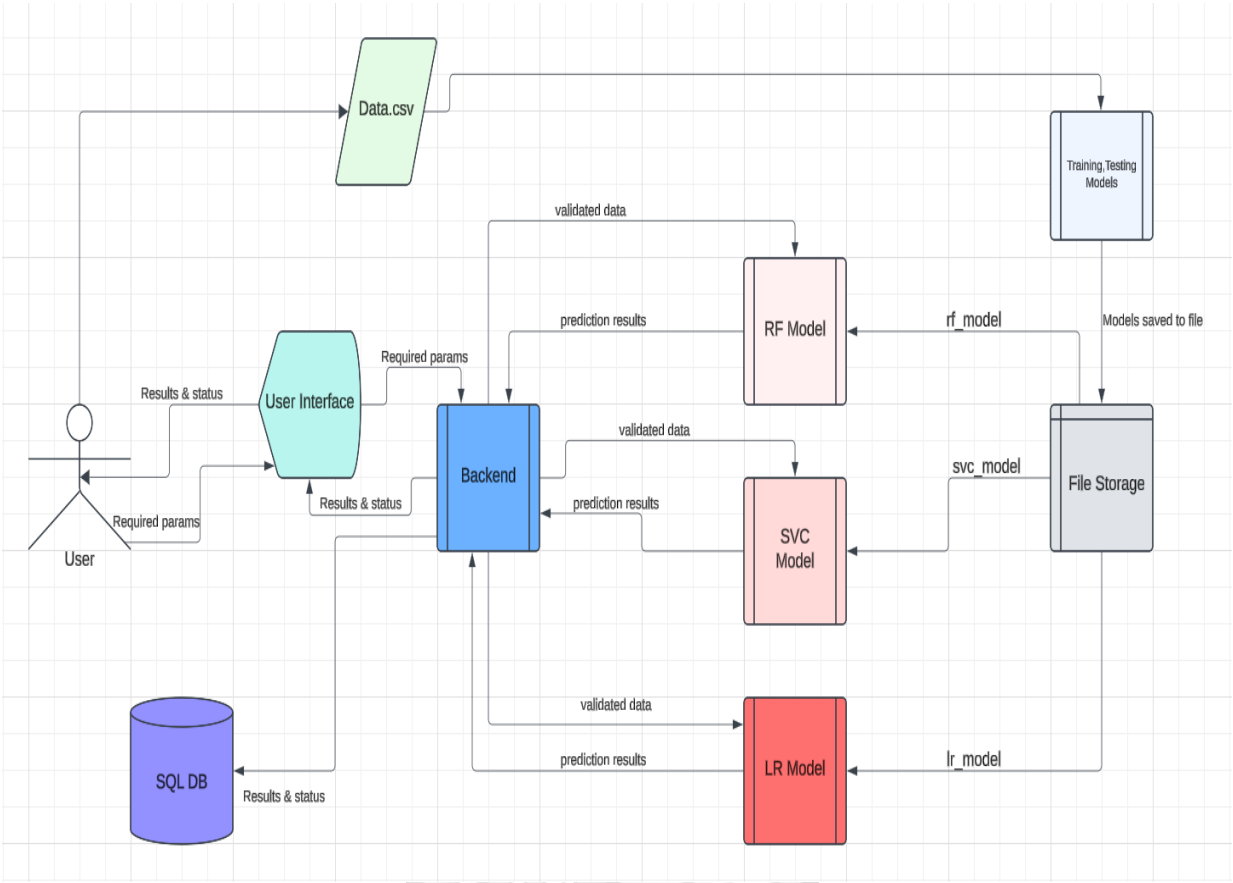


Figure 2.6: Conceptual Model of the solution.

Chapter 3: Research Methodology

3.1 Introduction

The growing frequency and intensity of flash floods in urban areas such as Nairobi necessitated the development of reliable prediction model to limit their negative consequences. The goal of this study was to create and apply a predictive model that is suited to Nairobi's specific hydrological and meteorological circumstances. The methodology used in this study was designed to ensure a thorough understanding of the components that contribute to flash floods and the development of a dependable prediction framework. The study technique was separated into major phases that is, research design, identifying target population, creating a sample size, data collection, preprocessing, model creation, and finally validation. Each phase was precisely constructed to handle the unique issues associated with flash flood prediction in Nairobi.

According to Igrewnagu (2016), research methodology is the systematic, theoretical analysis of the methods applied to a field of study. It encompasses the theoretical analysis of the body of methods and principles associate with a branch of knowledge. Kumar (1999) states that research is a way of thinking which examines critically the various aspect of day-to-day professional work, understanding and formulating guiding principles that govern a particular procedure, develops, and test new theories that contribute to the advancement of a field or profession. The adoption of choice regarding research design, data collection, and the target population provided the research with legitimacy and perspective on the study's boundaries. Methodology used by researcher supported the validity of the study (Somkh & Lewin, 2005).

3.2 Research Design

Research design is a procedural plan that is adopted by the research to answer questions validly, objectively, accurately and economically (Kumar, 1999). In their study, Kamiri and Mariga (2021) discovered machine learning research works mostly use quantitative analysis in combination with experimental research designs. Therefore, quantitative and experimental research methods was used this study as well to align with what other scholars use in the same field. Quantitative in the sense that there was calculation of data through machine learning algorithms and experimental since the data was divided into testing and training then the created model was evaluated with the testing dataset to validate its accuracy.

3.3 Target Population

The area of study for this research was Nairobi city as depicted in figure 3.1. Data used was secondary data related to climatic and topographic conditions of the city.

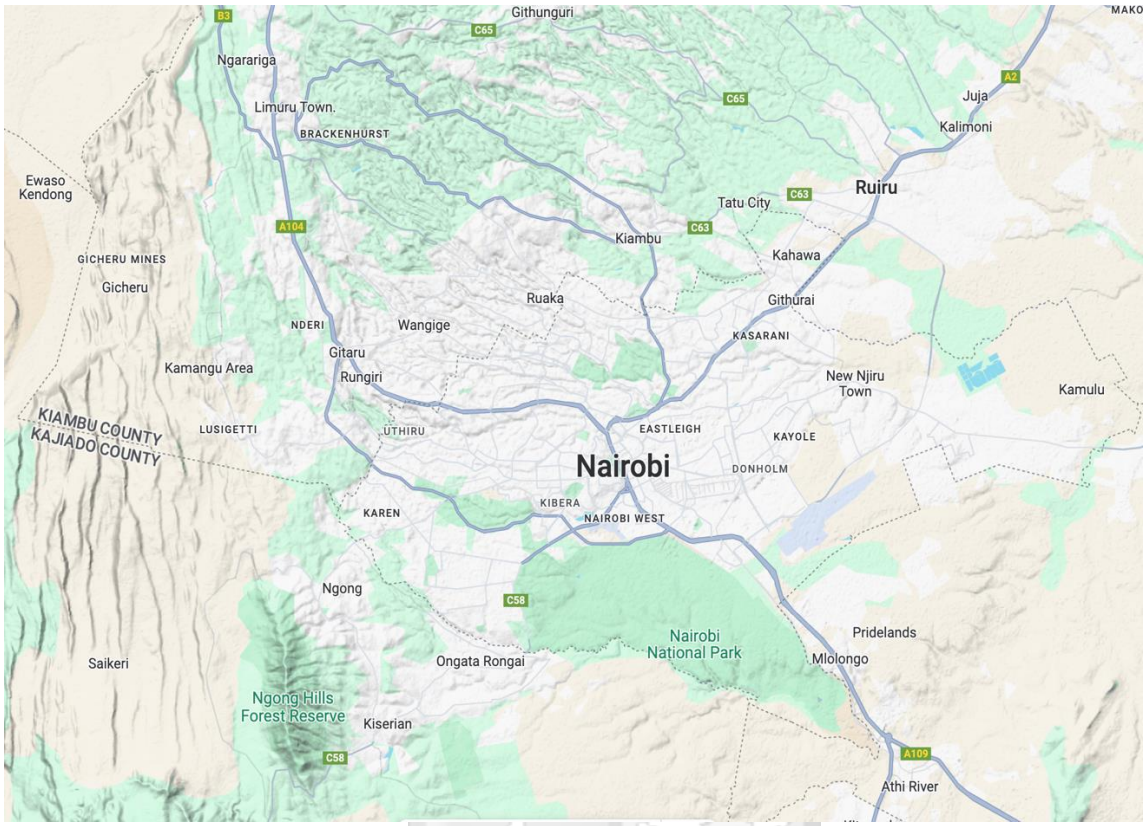


Figure 3.1: Map of Nairobi and satellite towns.

3.4 Sample Size

A sample size of 10% to 30% is needed for a descriptive survey (Mugenda & Mugenda, 2003). Rainfall and soil dataset from 2000 to 2022 was used as sample as data sample size to carry out this study and it was split for training and testing portions. In his article Brownlee (2020) said there is no optimal percentage in splitting the training and test datasets for machine learning however, the common split percentages are 70/30, 67/33, and 50/50. To avoid the model being biased and creating the false impression of a higher accuracy model, 30% of the dataset was utilized for testing and the remaining 70% for training. Because it strikes a suitable balance between training the model and evaluating its efficacy using the data at hand, the 70/30 rule is frequently used in machine learning a suggestion supported by Dobbin and Simon (2011) that a model may be trained on a large enough sample to identify patterns in the data and provide an estimate of its generalization performance on fresh data by designating 70% of the data as the training set and 30% as the testing set.

3.5 Data Collection

The dataset for this study was obtained from NASA's Prediction of Worldwide Energy Resources (POWER) project, accessible via the POWER Data Access Viewer (NASA, 2024). This platform provided a comprehensive collection of climate data, which was crucial for the development of our machine learning model. Specifically, the dataset included historical climate variables that is temperature, rainfall, soil moisture, humidity, and wind speed, spanning from the year 2000 to 2022. These climate variables were used to analyse patterns and forecast flash floods in the Nairobi region. In addition to climate data, the study incorporated soil topographical data, which was collected from local geological and hydrological studies. Table 3.1 and 3.2 show all the variables that were used in the analysis and model development.

Table 3.1: Climate Parameters

Description	**Units**
Temperature at 2 Meters Maximum	°C
Wind Speed at 50 Meters Maximum	m/s
Humidity	%
Rainfall	mm
Soil Moisture	%
River Discharge	M3s

Table 3.2: Soil Topography Parameters

Parameter	**Description**	**Values/Options**
Slope Position	Topographic position based on elevation	H: Hilltop or Upper slope, D: Depression or Valley, M: Mid-slope or Middle position, L: Lower slope
Surface Stoniness	Presence of significant surface stoniness	N: None (No significant surface stoniness)
Area Affected	The proportion of the area affected by erosion or deposition	Numerical value (e.g., "1" for 1% area affected, "4" for 40% area affected)
Erosion Degree	Degree of erosion severity	S: Slight erosion, V: Very severe erosion
Sensitivity to Capping	Sensitivity of soil to capping (hard crust formation on the soil surface)	N: None (No sensitivity to soil capping)
Land use type	Classification of the urban areas based on various categories	Agriculture, forest, urban, wetland
Flood occurrence	The overflow of water	1 for flood 0 Otherwise

3.6 Data Pre-Processing

Pre-processing is an essential step in data validation. The purpose of data pre-processing was to convert the data into a suitable form that algorithms can use since the algorithms only read data which are in numerical forms. Three main preprocessing steps that were applied to the dataset are data cleaning, feature encoding, and feature scaling. The pre-processing was done using Python programming language.

3.6.1 Data Cleaning

In the real world, data is often noisy and unstructured, making data cleaning a crucial step in ensuring the accuracy and reliability of any predictive model. This process involved identifying and addressing anomalies, such as missing values and outliers, to enhance the dataset's quality. In this study, data cleaning focused on handling missing values and smoothing out noise to ensure consistency in the dataset. The completeness of data allowed for seamless data preprocessing, ensuring that the model's predictions were based on a robust and well-structured dataset.

3.6.2 Features Encoding

In ML models, all inputs and outputs are required to be numerical variables. Therefore, when there are categorical data, it must be encoded before using it in the model, and this is called features encoding (Breskuvienė & Dzemyda, 2023). Feature encoding is the pre-processing of the categorical data when working on a model of ML algorithms. There are several techniques to encode categorical features, such as Label Encoding, One-Hot Encoding, Frequency Encoding, Ordinal Encoding, Binary Encoding, Hash Encoding, and Mean/Target Encoding. In this study's dataset, there were five categorical features, namely slope position, surface stoniness, erosion degree, sensitivity to capping, and land use type. To encode these features into numeric features, Label Encoder was employed because it is simple, efficient and maintains the order of categorical values.

3.6.3 Features Scaling

Feature scaling is a crucial preprocessing step that normalizes independent variables to ensure consistency across different scales, enhancing the performance of machine learning algorithms. By transforming data into a smaller range, typically between 0.0 and 1.0, feature scaling helps reduce error rates and accelerate the training process by preventing models from being biased toward larger numerical values. Several techniques exist for scaling data, including MinMaxScaler, StandardScaler, MaxAbsScaler, and RobustScaler, each with its

unique approach to handling different types of distributions. In this study, the MinMaxScaler technique was applied to the dataset, ensuring that all features were rescaled proportionally, thereby improving the efficiency and accuracy of the predictive model.

3.7 Splitting Dataset

The train-test split process was used. The data was first divided into features (X) and labels (y). The data frame was further split into training and testing sets of 70% and 30% respectively. X_train, X_test, y_train, and y_test. The X_train and y_train sets were used to train and fit the model, while the X_test and y_test sets were utilized to evaluate the model's performance in predicting the correct labels. Additionally, the sizes of the train and test sets were explicitly tested, with a recommendation to keep the training sets larger than the test sets. This process can be viewed on this GitHub repository Flash Flood Prediction Model repository.

3.8 Data Analysis

The study employed a classification-based approach to predict the occurrence of flash floods by analysing historical climate data and soil characteristics spanning from the year 2000 to 2022. The primary objective was to develop a predictive model capable of estimating the probability of floods occurring in specific areas. To achieve this, the study integrated multiple climate variables, including temperature, wind speed, and rainfall, alongside essential soil and terrain features such as erosion degree, area affected, and sensitivity to capping. These factors were chosen based on their established influence on flood susceptibility, ensuring that the model could effectively capture key patterns that contribute to flash flood events. By leveraging this diverse dataset, the study aimed at building a robust and reliable model that can aid in early warning systems and disaster preparedness initiatives.

To develop the predictive model, the study employed a variety of classification algorithms, including Linear Regression, Random Forest, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). Each of these algorithms was selected based on its ability to handle complex, high-dimensional data and identify meaningful relationships between independent variables and flood occurrence. Linear Regression, often used for establishing baseline models, helped in understanding the linear relationships among the variables. Random Forests, known for their ensemble learning capabilities, contributed to improved accuracy and robustness by aggregating multiple decision trees. KNN, a distance-based algorithm, was employed to assess local similarities in the data, while Artificial Neural Networks introduced deep learning techniques to capture nonlinear

patterns in flood susceptibility. Lastly, Support Vector Machines were utilized to maximize classification margins, ensuring an optimal separation between flood-prone and non-flood-prone areas.

The entire analysis was conducted using Python 3.9 and a suite of its powerful libraries, including NumPy, Pandas, Scikit-learn, TensorFlow, Matplotlib, and Seaborn. Data preprocessing, cleaning, and feature engineering were performed using Pandas and NumPy, ensuring the dataset was well-structured and free from inconsistencies. Machine learning models were implemented and trained using Scikit-learn and TensorFlow, providing a flexible framework for model selection, hyperparameter tuning, and optimization. The visualization of results, including feature importance and model performance metrics, was accomplished using Matplotlib and Seaborn, enabling a clear interpretation of the data. The coding and execution of the analysis were carried out in Jupyter Notebook**,** which facilitated an interactive and organized approach to data exploration, model training, and performance evaluation.

By adopting a classification-based methodology, this study provided actionable insights that could significantly enhance flood preparedness and response strategies. The predictive model offered a systematic approach to identifying factors that have a greater influence on flash flood occurrences, enabling policymakers, environmental planners, and disaster response teams to implement timely interventions. With the ability to assess flood risks based on historical climate patterns and terrain characteristics, decision-makers can proactively develop flood mitigation measures, improve early warning systems, and allocate resources more effectively to areas with the highest vulnerability. Ultimately, this research contributes to a data-driven approach in disaster management, ensuring a more resilient and prepared response to flash floods in the future.

3.8.1 Support Vector Machines (SVM)

Support Vector Machine (SVM) is a widely used supervised machine learning algorithm suited for classification tasks (Bithari et al., 2020). Introduced by Vapnick in 1995, SVM addresses prediction and pattern recognition challenges while analysing and mapping both linear and non-linear functions (Adejo & Connolly, 2018). The core principle of the SVM algorithm is to identify a hyperplane that maximally separates data into two distinct classes (Alboaneen et al., 2022). SVM achieves this by constructing one or more hyperplanes in a high-dimensional space, where objects are classified based on whether they lie above or below the plane, depending on their features. Additionally, kernel methods allow the algorithm to transform nonlinear relationships into linear ones before applying the partition.

3.8.2 Random Forest (RF)

Random Forest (RF) is a supervised machine-learning algorithm introduced by Leo Breiman in 2001 (Speiser et al., 2019). It operates by combining multiple decision trees to perform tasks such as regression and classification. The more trees within the forest, the stronger and more reliable the prediction becomes. Over time, RF has become a widely used and effective tool for data analysis and prediction (Alboaneen et al., 2022). In the case of regression, each decision tree generates a numerical prediction, and the Random Forest algorithm outputs the result by averaging these individual predictions.

3.8.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm, first introduced in the early 1950s, is a fundamental supervised machine learning technique used for both classification and regression tasks. It is widely recognized for its simplicity, effectiveness, and non-parametric nature, making it a valuable tool for pattern recognition and predictive modeling. Unlike many other machine learning algorithms that build an explicit model based on training data, KNN operates based on instance-based learning, meaning it does not construct an internal model during training. Instead, it memorizes the training data and defers computation until a prediction is required. This characteristic classifies KNN as a lazy learning algorithm, as it does not perform generalization until the classification or regression process begins (Han, Pei, & Tong, 2022).
The K-Nearest Neighbors (KNN) algorithm predicts the label of a new data point by identifying the K most similar instances from the training dataset using a selected distance metric, such as Euclidean distance, Manhattan distance, or Minkowski distance. In a classification problem, KNN assigns the majority class label among the K-nearest neighbors to the new instance. Because KNN follows a lazy learning approach, it does not build an explicit model during training. Instead, it stores the entire dataset and performs computations only when making predictions. This characteristic results in low training time but high prediction time, as every new prediction requires scanning the entire dataset. Consequently, KNN can be computationally expensive for large datasets, particularly when using a brute-force approach to find nearest neighbors. Despite its computational challenges, KNN is simple to implement and interpret, requiring minimal parameter tuning. It is particularly effective for small datasets and can easily be applied to multi-class classification problems, making it a versatile machine learning approach.

3.8.4 Artificial Neural Network (ANN)

McCulloch's research on simulating biological nervous systems in the 1940s laid the foundation for the development of Artificial Neural Networks (ANNs) (Alboaneen et al., 2022). ANNs are computational models designed to mimic the way the human brain processes information, learns from data, and adapts to changing environments. A Neural Network (NN) consists of multiple interconnected units (neurons) that process input data and generate outputs through weighted connections. The behaviour of an NN is influenced by its architecture (layers and connections), activation functions, learning algorithms, and training data. Over the years, ANNs have evolved into one of the most powerful and widely used machine learning techniques, enabling breakthroughs in image recognition, natural language processing (NLP), speech recognition, healthcare, finance, and robotics (Sharma et al., 2023). Artificial Neural Networks typically consist of three main layers: the input layer, the hidden layers, and the output layer. The input layer receives raw data, which is then processed through one or more hidden layers where mathematical transformations occur using activation functions like ReLU (Rectified Linear Unit), Sigmoid, or Tanh. Each connection between neurons has an associated weight, which determines the strength of the connection. During training, the model adjusts these weights using optimization techniques such as gradient descent and backpropagation, minimizing errors and improving accuracy. Finally, the output layer produces the final prediction or classification based on the learned patterns.

ANN was chosen in this study because of their ability to learn from data and recognize complex patterns without being explicitly programmed. Unlike traditional algorithms that require manually defined rules, neural networks automatically extract features from raw data, making them highly effective for tasks like image and speech recognition. Additionally, ANNs are capable of handling large volumes of high-dimensional data, making them suitable for real-world applications where massive datasets are involved. Their adaptability allows them to continuously improve and refine their performance over time, much like human learning.

3.8.5 Linear Regression (LR)

Regression methods are fundamental statistical and machine learning techniques used to model and analyse the relationship between a response variable (dependent variable) and one or more explanatory variables (independent variables). By establishing this relationship, regression analysis allows researchers to understand trends, quantify impacts, and make predictions based on historical data. Among the various regression techniques, Linear Regression (LR) is one of the simplest yet most powerful approaches, particularly within

supervised learning, where the goal is to predict continuous outcomes (Hope, 2020). Linear Regression assumes a linear relationship between the dependent variable and independent variables. When only one independent variable is used for prediction, the method is referred to as Simple Linear Regression. However, in practical applications, multiple factors often influence an outcome, necessitating the use of Multiple Linear Regression (MLR), where two or more explanatory variables are included (Dash et al., 2023).

Linear Regression was selected as a key predictive method for flash floods in Nairobi County due to its ability to quantify the impact of various climatic and environmental factors on flood occurrence. The region experiences seasonal rainfall variability, rapid urbanization, and soil degradation, all of which contribute to flash flooding. By using LR, the study aimed to identify and quantify key predictors such as rainfall intensity and frequency, soil moisture levels, land surface characteristics, temperature variations, wind speed, and topographical factors. One of the primary advantages of using Linear Regression for flash flood prediction is its interpretability and simplicity. LR provides a clear and interpretable model where each coefficient quantifies the effect of an independent variable on the likelihood of flash floods. Unlike complex machine learning models, LR allows researchers to directly analyse the influence of rainfall, temperature, and land characteristics on flood probability. Additionally, Linear Regression is computationally efficient and requires fewer resources compared to non-linear models like Artificial Neural Networks (ANNs) or Random Forests. This makes it ideal for large-scale climate datasets, allowing for faster analysis and real-time predictions.

Furthermore, LR is particularly effective for continuous flood probability estimation. Since flash flood occurrence is highly correlated with meteorological and geographical conditions, LR helps estimate the probability and severity of flooding based on historical data. This enables authorities to predict future flood risks based on weather forecasts and urban development patterns. Another key benefit is its ability to identify the most critical flood predictors, allowing policymakers to focus on the most influential factors. For instance, if the model reveals that urbanization (impervious surfaces) significantly increases flood risk, city planners can implement better drainage systems and sustainable urban development strategies.

3.9 Model Evaluation

The model was rigorously evaluated using performance metrics of accuracy, precision, recall, and F1-score to assess its predictive reliability. These metrics provided a comprehensive evaluation of the model's effectiveness in correctly identifying flood-prone while minimizing false predictions. Accuracy measured the overall correctness of predictions, while precision

and recall provided insights into the trade-off between false positives and false negatives, ensuring the model was balanced and reliable. The F1-score, a harmonic mean of precision and recall, further ensured that both false alarms and missed flood predictions were minimized, making the model suitable for real-world applications.

The formula below shows how accuracy, precision, recall, and F1 will be calculated.

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$F1 = 2 * ((\text{Precision} * \text{Recall})/ (\text{Precision} + \text{Recall}))$$

Where TP refers to True Positive, TN refers to True Negative, FP refers to False Positive, and FN refers to False Negative. These were derived from the confusion matrix, which assists in classifying a model's performance, as shown in Table 3.3.

Table 3.3: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Positive	TN	FP
Actual Negative	FN	TP

3.10 Research Result's Utilisation and Dissemination

Flash flood control, prediction, and mitigation are critical in any society or economy. This study aimed to help Nairobi residents, county governments, and humanitarian organizations prepare for potential flash flood disasters and take proactive measures to prevent loss of life and minimize economic damage. This will help to reduce catastrophes and losses that occur whenever there is a flash flood or abnormal weather conditions in the city. The results and findings of this study will be shared online through the Strathmore University repository and easily accessed by anyone. The created model will be hosted on either Heroku or Digital Ocean where anyone could use the application interfacing the model to make predictions. In addition, the codebase will be hosted on GitHub where those interested in learning how it works could clone it and make suggestions and contribute to its growth.

3.11 Ethical Considerations

The study underwent the mandatory Institutional Scientific and Ethical Review Process to ensure compliance with established research guidelines and ethical considerations. Ethical approval was obtained from Strathmore University, affirming the study's adherence to rigorous

academic and scientific standards as indicated on appendix B while originality was determine using Turnitin as indicated on appendix A. Additionally, a research license was secured from the National Commission for Science, Technology, and Innovation (NACOSTI), further validating the study's credibility and authorization to conduct research within Kenya as indicated on appendix C.

The data utilized in this study was collected exclusively from publicly accessible and reputable platform, NASA's Prediction of Worldwide Energy Resources (POWER) data viewer, ensuring full compliance with legal and privacy regulations governing research. To uphold scientific integrity and transparency, proper attribution was provided for all data sources, acknowledging their contributions to the study. The study prioritized data security by implementing strict access controls. The collected data was securely stored, with restricted access granted only to authorized personnel directly involved in the study. This measure prevented unauthorized use and ensured that the data was utilized solely for the intended research objectives. Additionally, transparency was maintained throughout the study regarding the purpose of data collection, aligning with ethical research guidelines and promoting accountability. Therefore, by adhering to these ethical standards, obtaining the necessary approvals, and ensuring data integrity, the study upheld the principles of responsible research while contributing valuable insights to flood risk prediction and mitigation efforts

Chapter 4: System Analysis and Design

4.1 Introduction

System Analysis and Design (SAD) is a systematic methodology to develop and maintain information systems that align with organizational objectives (GeeksforGeeks, 2024). The system analysis phase involves a comprehensive study of the current system to understand its components, workflows, and interactions. This analysis helps identify problems, inefficiencies, and opportunities for improvement. Data flow diagrams are commonly used to gather relevant information from the system. Therefore, this section describes the created model, process architecture, components, and key functionalities.

4.2 System Design

System design refers to the process of defining the architecture, components, modules, interfaces, and data flow of a system to meet specific functional and non-functional requirements. It provides a structured blueprint that outlines how different elements of a system interact and work together to achieve the desired objectives. In the context of this study, system design involves planning how data will be collected, processed, analysed, and used to make flash floods predictions. It ensures that the flash floods system is efficient, scalable, and capable of delivering accurate flood forecasts in a timely manner.

4.2.1 Functional Requirement

Functional requirements describe the specific behaviour or functions of a system. In this case the requirements describe how flash floods system responds to inputs, reactions, and behaviour under certain conditions, and output (Requirements.com, 2024). The requirements were derived from user needs and significant in transforming user expectations into technical specifications. In this study, the functional requirements were:

i. The flash flood system should allow users to upload file with climatic data to be used for prediction.
ii. The system should be able to validate and process flash flood data for consistency and quality, which is crucial for reliable model performance.
iii. The system should be able to leverage the machine algorithms to train models.
iv. The system should use the trained models to predict flash flood occurrence and produce output.

4.2.2 Non-functional Requirement

Non-functional requirements specify how the system performs its tasks. It focuses on the attributes of the system, such as performance, security, scalability, and usability, which determine how well the system performs its functions (GeeksforGeeks, 2025). Non-functional requirements for this study were:

i. Performance - the system should process data and generate predictions within a reasonable time frame of 3 seconds.
ii. Scalability - the system should be able to handle increased data volume and simultaneous user numbers, especially during peak demand. It should also be modular to allow for addition of features when necessary.
iii. Reliability - the system should be able to maintain continuous system operation with a failover mechanism to ensure reliability during disasters.
iv. Accuracy - the system must achieve high prediction accuracy to minimize false alarms and missed events.
v. Ease of Use - the system should provide an intuitive interface that non-technical users can easily understand and navigate.

4.3 System Architecture

System architecture defines the structural design and interactions between various system components to ensure their functionality, scalability, and efficiency (Carter & Singam, 2024). It includes identifying software components, their relationships, interfaces, and data flows. The flash flood prediction system in figure 4.1 was designed with a modular architecture to ensure efficient data handling, prediction accuracy, and user-friendly and intuitive user interface. The system was designed as per the objectives of the study. First hydrometeorological data is collected and saved into a file of '.csv' format which was then uploaded into the system by a user. It is then pre-processed and cleaned using Python programming language libraries in this case Pandas, NumPy, and Scikit-learn. Once cleaned, it is used to train models using ML algorithms as mentioned in previous chapter. The trained models are saved into a file where they can be utilized by any user to make real-time flash flood predictions. The best performing models were selected with the criteria for selection of the best performing model being an accuracy result of 90% and above. There is an application with a user interface which sits between the user and the models. This was done to ensure that the system was user friendly,

and anyone can use it to make prediction. User is required to input the climatic conditions identified in chapter three into a web form and click on the submit button. The process of flash flood prediction will be initiated where the keyed in features will be validated and categorized by the application. Once the features are validated, the application will feed these features to the three best performing models to make prediction. Once the predictions are made, the results will be fed back to the application which will get the average of the results and save it to the database and relay the result to the user.

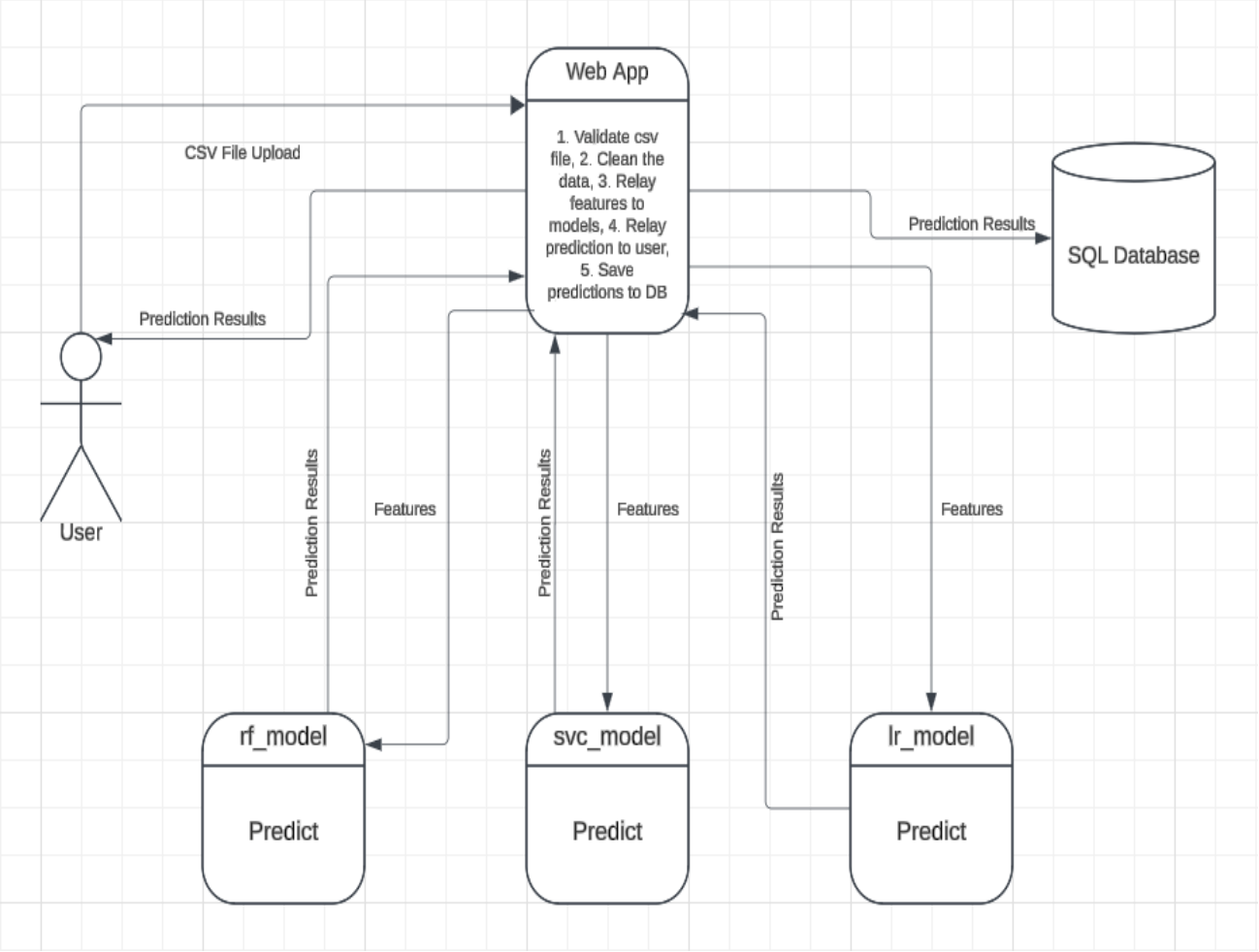


Figure 4.1: System Architecture

4.4 Use Case Diagram

A use case diagram is a graphic depiction that highlights the various functionalities offered by a system and shows how users interact with it. The use case diagram outlines how several personas, including meteorologists, disaster management organizations, local authorities, and the public, engage with the early warning system in the context of the Nairobi County Flash Flood Prediction Model. The key use cases included data input and processing, which collects and pre-processes real-time climate variables and soil moisture index data;

model execution, which analyses data using supervised machine learning algorithms to predict flash flood risks and severity levels; and report visualization, which displays results on interactive geospatial dashboards. The use case diagram clarifies the system's operation, assures a user-centric flood risk communication design, and indicates future improvements for smooth integration into community-based early warning systems and disaster response frameworks as depicted in Figure 4.2.

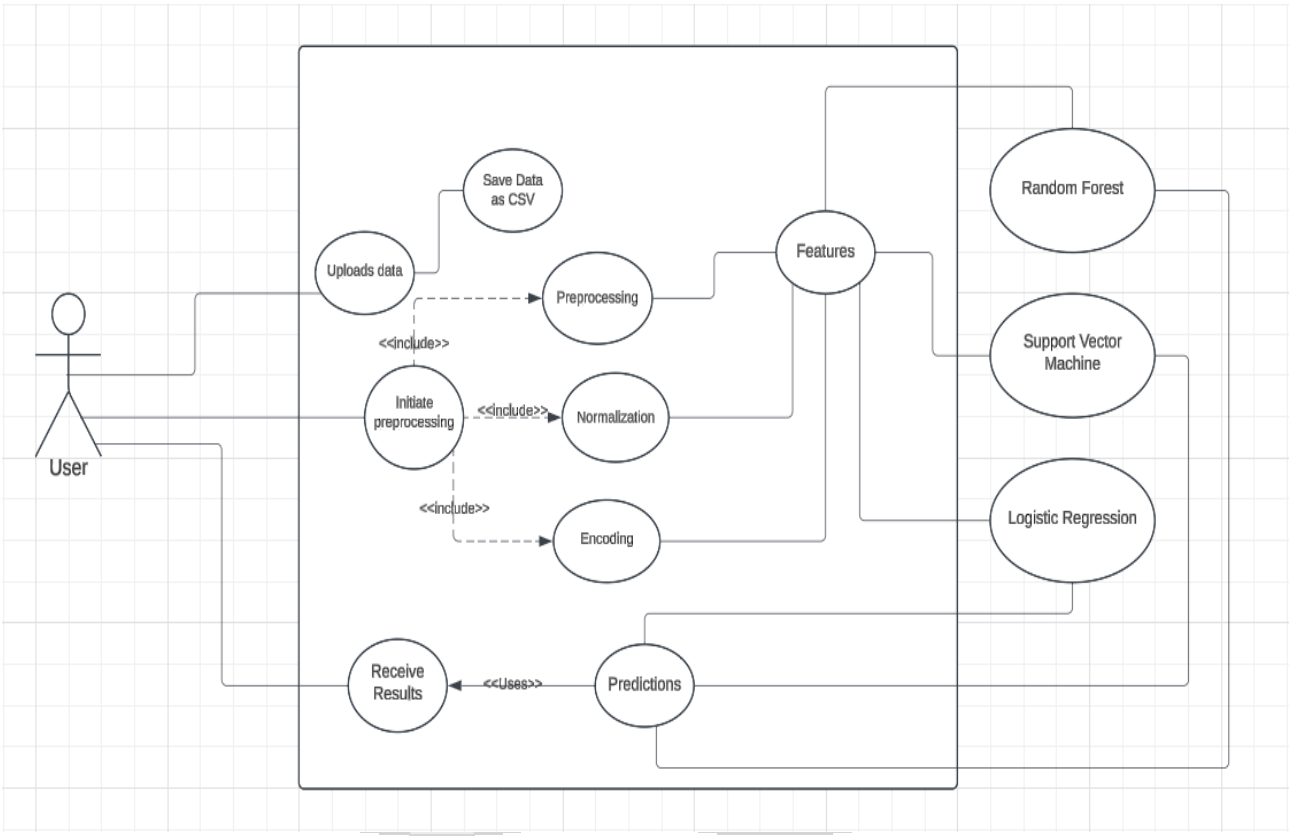


Figure 4.2:Use Case Diagram

The primary actors in the use case diagram are the users and the system itself. The use cases for both the primary actors are highlighted below:

Use case: Upload Data.

i. Primary Actors: User.

ii. Pre-conditions: The user can access the internet on the platform/device being used.

Use case: Initiate Program

i. Primary Actors: User, System.

ii. Pre-conditions: Retrieve use case completed successfully.

Use case: Pre-Process

i. Primary Actors: User, System.
ii. Pre-conditions: Retrieve use case completed successfully.

Use case: Predict Flash Floods

i. Primary Actors: System.
ii. Pre-conditions: Transform use case completed successfully.

Use case: Receive Results

i. Primary Actors: User.
ii. Pre-conditions: Predict flash floods use case completed successfully.

4.5 Sequence Diagram

A sequence diagram is an interaction diagram in Unified Modeling Language (UML) that illustrates how objects or components within a system interact over time (GeeksforGeeks, 2025). It visually represents the sequence of messages exchanged between participants, highlighting the order and timing of these interactions in a system. Sequence diagrams particularly useful for modeling the flow of control in a system, detailing the specific steps involved in a process or use case. Sequence diagrams help understand a system's dynamic behaviour by focusing on the temporal aspect of interactions, which is essential for system design and analysis. Figure 4.3 was the sequence diagram of message flows in the model created by this study.

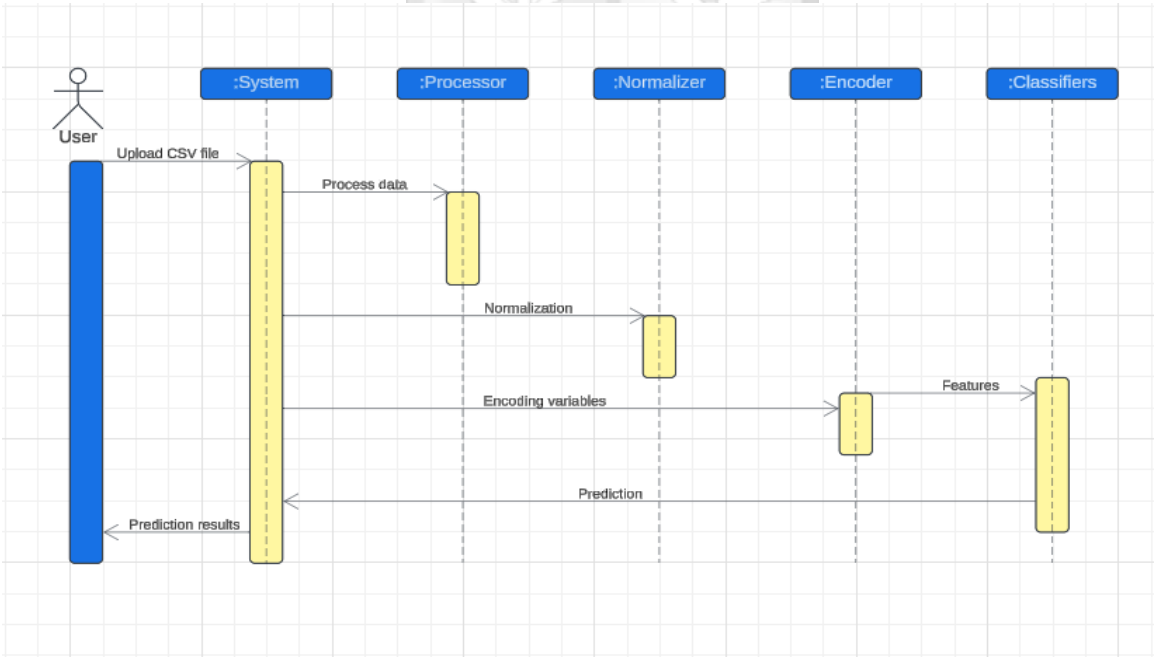


Figure 4.3: Sequence Diagram

Chapter 5: Model Implementation and Testing

5.1 Introduction

The study was guided by three key research questions: i. What are the factors that influence the occurrence of flash floods in Nairobi County? ii. How can a machine learning model be developed to predict flash floods accurately? iii. How effective is the developed model in predicting flash floods? Addressing these research questions, this chapter contributed to the broader understanding of how advanced computational methods can enhance flash flood prediction, informing disaster preparedness and mitigation efforts in urban areas like Nairobi County. This study's theoretical framework also integrated concepts from hydrology, meteorology, and data science, emphasizing the interplay between environmental variables and machine learning techniques. Specifically, the framework drew on predictive analytics and disaster risk management principles, allowing for a structured exploration of flash flood predictors and model development.

5.2 System Implementation

System implementation refers to the process of deploying the developed Flash Flood Prediction Model for Nairobi County into a functional environment where it can be used for real-time predictions and decision-making. This phase involved setting up the necessary computing infrastructure, integrating machine learning models, and ensuring smooth data processing. The implementation began with data acquisition and preprocessing, where historical and real-time climate and soil data were collected from NASA's POWER Data Viewer database. The machine learning algorithms used were Linear Regression, Random Forest, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). The models were trained, tested, and optimized using Python 3.9 programming language and its data manipulation libraries to ensure high predictive accuracy.

Once the models were validated, the system was deployed on Heroku, a cloud-based platform to facilitate accessibility for stakeholders, including meteorologists, disaster response teams, and policymakers. A user interface was designed to display predictions through dashboard, providing real-time flood risk alerts. Additionally, continuous monitoring and maintenance are crucial to refine model accuracy, update data sources, and ensure system security, making the flood prediction model a reliable tool for disaster preparedness and mitigation.

5.2.1 Data Processing

The machine learning model development involved several essential steps, including data preprocessing. Some of the preprocessing steps which were conducted included checking missing values, scaling, and encoding. The dataset had no missing values, so the focus was on scaling and encoding the continuous variables. Scaling was applied to the continuous variables using the MinMaxScaler from scikit-learn. The features scaled were wind speed (km/h), temperature, soil moisture (%), humidity (%), rainfall (mm), river discharge (m^3/s), and elevation (m). The MinMaxScaler transformed these variables to a range between 0 and 1, ensuring the features were comparable. This step is crucial for many machine learning models, particularly distance-based algorithms like KNN and SVM, as it prevents any variable from disproportionately influencing the model's performance due to its more significant scale.

In addition to scaling, encoding was applied to categorical variables. The categorical features encoded were slope position, surface stoniness, erosion degree, sensitivity to capping, and land use type. Label encoding was used for these variables, transforming each unique category within the features into a corresponding integer value. This encoding process allowed the machine learning algorithms to efficiently process categorical data by converting it into a format they can understand and utilize in making predictions. The model leveraged all relevant information by appropriately handling these categorical variables. Figures 5.1 and 5.2 show the normalization and encoding processes, respectively.

Figure 5.1: Normalization of continuous variables

```
[59]: #Convert categorical features to numeric using LabelEncoder()

label_encoder = LabelEncoder()

# Converting 'Slope_Position' data into numerical values
label_encoder.fit(data["Slope_Position"])
slope_data_encoded =  label_encoder.transform(data['Slope_Position'])
data['Slope_Position'] = slope_data_encoded

# Converting Surface Stoniness into numeric values
label_encoder.fit(data['Surface_stoniness'])
stoniness_data_encoded = label_encoder.transform(data['Surface_stoniness'])
data['Surface_stoniness'] = stoniness_data_encoded

# Converting Erosion Degree into numeric values
label_encoder.fit(data['Erosion_degree'])
erosion_degree_data_encoded = label_encoder.transform(data['Erosion_degree'])
data['Erosion_degree'] = erosion_degree_data_encoded

# Converting 'Sensitivity to capping' into numeric values
label_encoder.fit(data['Sensitivity_to_capping'])
sensitivity_to_capping_data_encoded = label_encoder.transform(data['Sensitivity_to_capping'])
data['Sensitivity_to_capping'] = sensitivity_to_capping_data_encoded

# Converting 'Land Use Type' into numeric values
label_encoder.fit(data['Land_Use_Type'])
land_use_type_data_encoded = label_encoder.transform(data['Land_Use_Type'])
data['Land_Use_Type'] = land_use_type_data_encoded
```

Figure 5.2: Converting categorical variables to numeric

5.2.2 Data Splitting

Following preprocessing, the dataset was split into training and testing sets, ensuring an even distribution of data and enabling the evaluation of the model's performance on unseen data. A 70-30 split was typically used, where 70% of the data was used for training while 30% was used for testing. Figure 5.3 shows the data splitting process. The process was important as it heled to minimize the risk of overfitting and provided a more accurate estimate of the model's generalization ability.

```
Split data into training and testing sets

[80]: X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
      print("XTrains->",X_train.shape[0],"XTest->",X_test.shape[0],"YTrain->",y_test.shape[0],"YTest->",y_test.shape[0])

      XTrains-> 140 XTest-> 60 YTrain-> 60 YTest-> 60
```

Figure 5.3: Data splitting

5.3 Feature Selection

Feature selection was a crucial step in developing the flash flood prediction model for Nairobi County**,** as it determined the most relevant variables that significantly impacted flood occurrence. By selecting the most influential features, the model improves accuracy, reduces computational complexity, and minimizes the risk of overfitting. In this study, the correlation matrix was used to determine features that influence flash floods as shown in figure 5.4. From the matrix, rainfall had a strong positive correlation of 0.61 with flood occurrence, as increased rainfall is one of the primary causes of flash floods. Soil Moisture also showed a moderate positive correlation (0.52), indicating that higher soil moisture levels are associated with an

increased likelihood of flooding. This was consistent with the understanding that saturated soil is less able to absorb additional rainfall, increasing surface runoff and flood risk.

River discharge, with a correlation of 0.48, was another key factor influencing flash floods. This suggested that the flow rate of rivers is an essential determinant of flood occurrence. When river discharge is high, it can exceed the capacity of the riverbanks, leading to flooding. Erosion Degree (0.37) also played a significant role, as areas with higher erosion are more prone to losing soil integrity, resulting in increased surface runoff and flash floods. Land use type showed a relatively weak positive correlation (0.16) but was still relevant. Different land uses, such as urbanization or agricultural activities, can alter natural drainage patterns, potentially increasing the risk of flooding. Humidity and Temperature exhibited minimal positive correlations (0.09 and 0.08, respectively), indicating that they may have a secondary or indirect influence on flash flood occurrence compared to the other factors. Elevation (m) and surface stoniness had very weak positive correlations, 0.08 and 0.07, respectively, which suggested that while they might contribute to local variations in flood risks, their overall impact on flood occurrence was limited. Conversely, the area affected (0.06) and wind speed (0.04) showed even weaker positive correlations, suggesting that these factors play a minimal role in the occurrence of flash floods in Nairobi County.

Interestingly, the sensitivity to capping correlation was relatively weak at -0.05, and the slope position correlation was -0.16 with flood occurrence. This indicated that areas with lower sensitivity to capping or certain slope positions may be less prone to flash floods, possibly due to better drainage or more stable terrain that resists flooding. Therefore, the primary factors influencing flash flood occurrence in Nairobi County were rainfall, soil moisture, river discharge, and erosion degree, while land use type, humidity, and temperature contributed to a lesser extent. Environmental and geographical factors such as elevation, surface stoniness, and slope position also play a role, but their impact was relatively small compared to the more direct drivers of flooding. The four factors were selected and used to train the models. Figure 5.4 shows the correlation matrix of the factors that influence flood occurrence.

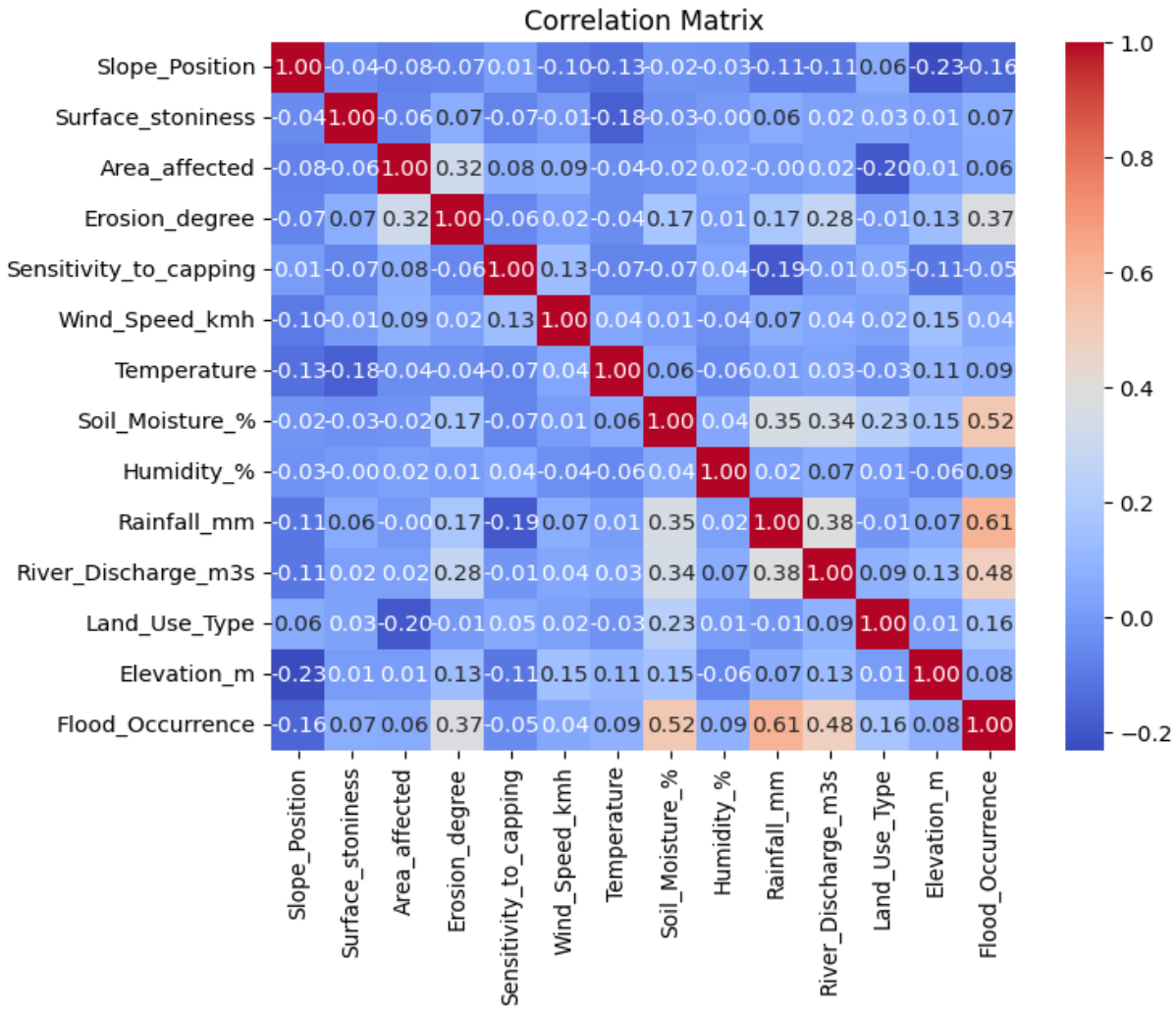


Figure 5.4: Correlation matrix

5.4 Model Training and Testing

Five machine learning models were trained, tested, and evaluated to determine the best-performing algorithm for predicting flash flood occurrence. Random Forest, K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were chosen for comparison. Each model was trained using the pre-processed training data, and their performance was evaluated using the testing set. Hyperparameter tuning was performed on each model to optimize their performance. For Random Forest, this involved adjusting the number of trees and the maximum depth of each tree, while for SVM, parameters like the kernel type and regularization strength were tuned. For KNN, the number of neighbors was adjusted, and for Logistic Regression, the regularization strength was optimized.

5.4.1 Random Forest

The Random Forest model was constructed using Scikit-Learn's RandomForestClassifier class module, which was imported and utilized within the model's class, as illustrated in Figure 5.5. From the figure, the max_depth=3 parameter specifies the maximum depth of each decision tree, limiting the number of levels to ensure the trees are not overly complex and help prevent overfitting. The random_state=0 parameter was used to ensure the reproducibility of the results by fixing the randomization process during bootstrapping and feature selection. During training, the model employed an ensemble learning approach where multiple decision trees were constructed. Each tree was trained on a random subset of the training dataset, created using bootstrapping, which involves sampling with replacement. At each split in the trees, features were randomly selected, and the best split was determined to improve the model's ability to handle variability and avoid overfitting. The fit function was used to train the Random Forest model on the training dataset. Each decision tree in the forest learned independently and contributed to the overall model's predictive power by voting for the most likely class. Predictions were then made on the test dataset using the predict(X_test) function. For every data point in X_test, the model passed the input features through each decision tree in the forest. The trees collectively voted, and the majority class prediction became the final output. This voting mechanism improves prediction accuracy and robustness by reducing variance and mitigating overfitting compared to a single decision tree. The result of this process, stored in y_pred, contains the predicted class labels for the test dataset. This output represents the model's attempt to classify each data point based on patterns learned during training.

Figure 5.5: Fitting and testing the Random Forest Classifier

5.4.2 Support Vector Machine

Figure 5.6 shows how Support Vector Machine (SVM) model was constructed using Scikit-Learn's SVC class module, configured with a linear kernel and a regularization parameter C=1.0. The linear kernel specified that the model separated the data points in the feature space using a straight-line hyperplane. The regularization parameter, C, balanced

achieving a low training error and maintaining a smooth decision boundary. A smaller value of C allowed for more margin violations, leading to a simpler and more generalized model, while a more significant value of C focuses on minimizing the training error. The fit function was used to train the model, where X_train represents the input features of the training dataset, and y_train contains the corresponding target labels. During training, the SVM algorithm identified the optimal hyperplane that maximizes the margin, the distance between the hyperplane, and the nearest data points, called support vectors. Maximizing this margin ensured the model's robustness and ability to separate classes effectively. Once the model was trained, the predict(X_test) function was applied to the test dataset, X_test, which contains unseen input features. For each data point in the test set, the model determined its position relative to the hyperplane and assigned it to one of the classes. The predicted class labels, stored in y_pred, represent the model's classification results based on the patterns learned during training. This process demonstrates how the SVM model effectively uses the hyperplane to classify data points in a linear feature space.

```
[173]: support_vector_model = SVC(kernel = kernels[0])
       support_vector_model.fit(X_train.values, y_train.values)
       svc_prediction_result = support_vector_model.predict(X_test.values)
       # print(f"Support Vector Prediction results {svc_prediction_result}")
       svc_accuracy = accuracy_score(y_test, y_pred) * 100
       svc_recall = recall_score(y_test,y_pred) * 100
       svc_f1 = f1_score(y_test, y_pred) * 100
       print(f"Accuracy {svc_accuracy:.2f}")
       print(f"Recall {svc_recall:.2f}")
       print(f"F1 Score {svc_f1:.2f}")
```

Figure 5.6: Support Vector Machine Classifier

5.4.3 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) model was developed using Scikit-Learn's class module, which was imported from the neighbors' package. This algorithm shown in Figure 5.7 is an effective supervised machine learning method that classifies data points based on the similarity of their features to neighboring data points. An instance of the KNN classifier was created, which initialized the classifier with default parameters. The default settings determine the number of neighbors (k=5) and the distance metric to identify the closest points. The classifier was then trained on the dataset using the fit (X_train, y_train) function, where X_train represented the input features of the training dataset, and y_train contained the corresponding target labels. During this process, the model stored the training data points, as KNN relies on the proximity of points for predictions rather than explicitly learning a parametric decision boundary. After training, the predict(X_test) function was used to classify the test data points in X_test, which contained previously unseen input features. For each data point in the test set, the algorithm calculated the distance to its nearest neighbors in the training data and assigned

a label based on the majority class among those neighbors. The predicted labels were stored in y_pred, representing the model's classification results. This approach allowed the KNN model to make predictions by leveraging the similarity of data points within the feature space.

```
[90]: # from sklearn import model_selection,neighbors
      classifier = neighbors.KNeighborsClassifier()
      knn_classifier = classifier.fit(X_train,y_train)

[91]: #Predict chances of flood
      y_predict = knn_classifier.predict(X_test)
```

Figure 5.7: K-Nearest Neighbors Classifier

5.4.4 Logistic Regression

The Logistic Regression model was also built using Scikit-Learn's class module. The model was then trained on the dataset using the fit (X_train, y_train) function, where X_train represented the input features of the training set, and y_train contained the corresponding target labels. During the training process, Logistic Regression used a linear combination of the input features and applied the sigmoid function to estimate the probability of each class, optimizing the parameters to minimize the difference between predicted and actual class labels. Once trained, the predict(X_test) function was used to make predictions on the unseen test dataset, X_test. The expected class labels were stored in y_pred, representing the model's classification output based on the learned relationships between features and target labels. Figure 5.8 illustrates the development of Logistic Regression.

```
[95]: logistic_regression_model = LogisticRegression()
      logistic_regression_classifier = logistic_regression_model.fit(X_train,y_train)
      logistic_regression_accuracy = cross_val_score(logistic_regression_classifier,X_test,y_test,cv=3,scoring='accuracy',n_jobs=-1)

[96]: y_predict = logistic_regression_classifier.predict(X_test)
```

Figure 5.8: Fitting and Testing the Logistic Regression Classifier

5.4.5 Artificial Neural Network

The code in Figure 5.9 demonstrates the implementation of a neural network for binary classification using TensorFlow's Keras library. The model was constructed with a sequential architecture, where the number of features in the training data was determined by X_train and shape, representing the number of input columns. The network consists of three layers: the first was a dense layer with 64 neurons using the ReLU activation function and an input size equal to the number of features. The second layer was another dense layer with 32 neurons, utilizing the ReLU activation function. The final layer was a dense output layer with one neuron and a sigmoid activation function, which was suitable for binary classification as it outputs

probabilities between 0 and 1. The model was compiled using the Adam optimizer, known for its efficient gradient descent performance, and the binary cross-entropy loss function, which was ideal for measuring the error in binary classification tasks. The accuracy metric was also included to track the model's performance during training and validation. Two callbacks were defined to enhance training efficiency and reliability. The EarlyStopping callback monitors the validation loss, stopping training if it does not improve for 10 consecutive epochs, and restores the weights from the epoch with the best validation loss. The ModelCheckpoint callback saves the best model to a file named best_model.h5, ensuring the optimal version is retained. The model was trained using the fit method, where X_train and y_train were the input features and labels for training, while X_test and y_test were used for validation. The training ran for a maximum of 100 epochs with a batch size of 32, but it was terminated early due to the EarlyStopping callback. The callbacks ensured the training process was efficient and prevented overfitting, with the best-performing model saved for future use.

```
ann_model = Sequential()
ann_model.add(Dense(units = 64, activation = 'relu', input_dim=X_train.shape[1]))
ann_model.add(Dense(units = 32, activation = 'relu'))
ann_model.add(Dense(units=1))

# Compile Model
ann_model.compile(optimizer='adam', loss='mean_squared_error')

# Train the model
history = ann_model.fit(X_train.values, y_train.values, epochs=100, batch_size=32, validation_split=0.2)

#Evaluate the model
loss = ann_model.evaluate(X_test.values, y_test.values)
print(f"Test Loss: {loss}")
y_predict = ann_model.predict(X_test.values)
```

Figure 5.9: Artificial Neural Network Classifier

5.5 Model Testing and Validation

Testing and validation of a model is a critical phase in developing a prediction model, ensuring that the model is accurate, reliable, and generalizable to real-world scenarios. This study evaluated the performance of each of the five models using accuracy, recall, precision, and F1-score. The Random Forest model and Support Vector Machine emerged as the best performers. Random Forest achieved an impressive accuracy of 93.33% with a recall score of 90.47% indicating the strong predictive power, particularly in correctly identifying flash flood occurrences. Furthermore, the F1 score of 0.90 highlighted a well-balanced model regarding precision and recall, making it a reliable choice for flood prediction. On the other hand, the Support Vector machine had a recall score of 85.71% and an F1 score of 0.9. The confusion matrix in Figure 5.10 provides a detailed breakdown of the model's performance in classifying instances into no flood occurrence and flood occurrence. The model accurately predicted no flood occurrence with 37 True Negatives. Similarly, it achieved many True Positives (19),

correctly identifying flood occurrence. On the other hand, it generated 2 False Positives, incorrectly predicting flood occurrence when there was no flood occurrence. Furthermore, 2 False Negatives were observed, indicating instances where the model failed to detect actual flood occurrence.

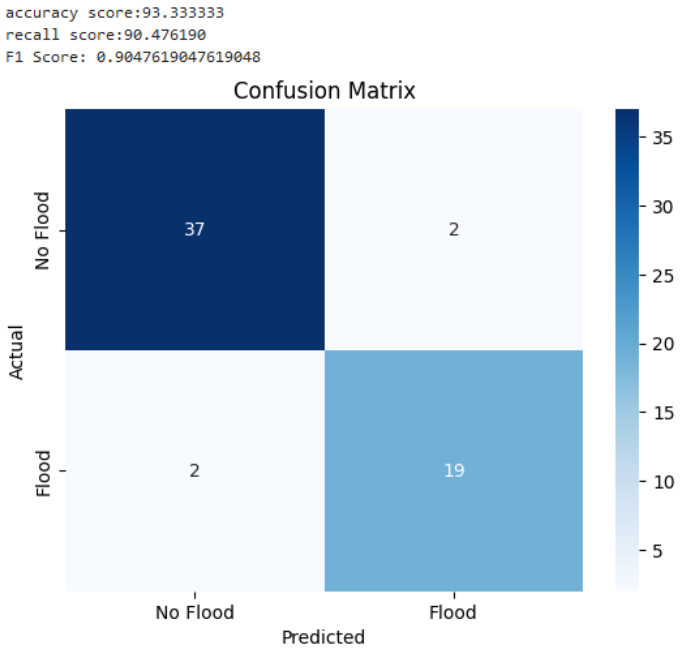


Figure 5.10: Confusion Matrix for Random Forest

Similarly, the confusion matrix in Figure 5.11 provides a detailed breakdown of the Support Vector Machine model's performance in classifying instances into no flood and flood occurrences. The model accurately predicted no flood occurrence with 38 True Negatives. It achieved many True Positives (18), correctly identifying flood occurrence. On the other hand, it generated 1 False Positive, incorrectly predicting flood occurrence when it was no flood occurrence. Furthermore, 3 False Negatives were observed, indicating instances where the model failed to detect actual flood occurrence.

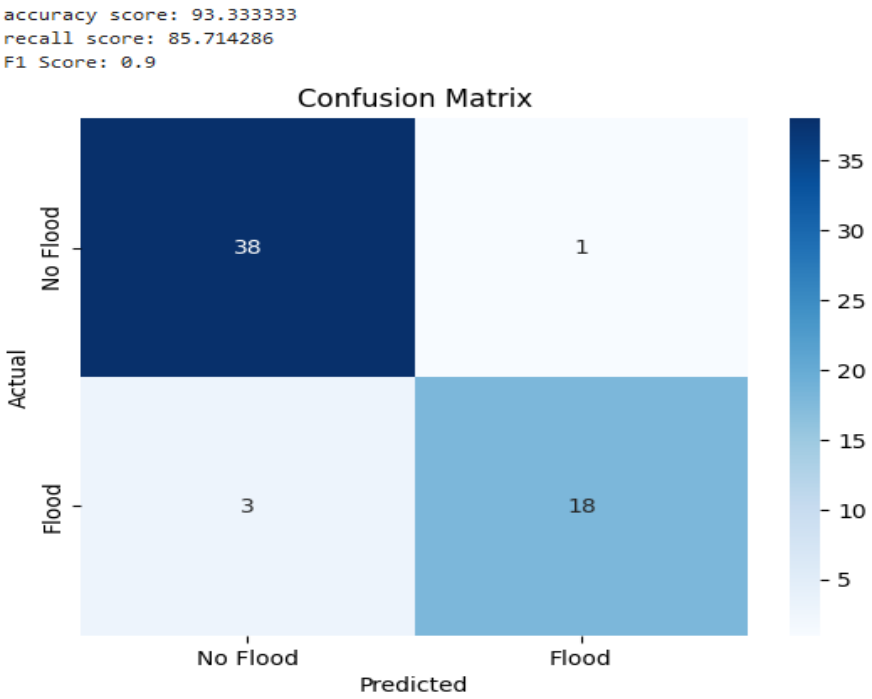


Figure 5.11:SVM Confusion Matrix

The K-Nearest Neighbors (KNN) algorithm demonstrated impressive predictive performance in the flash flood prediction. It achieved an accuracy of 86.67% as shown in Figure 5.12, which indicates that the model correctly classified flood occurrences in most cases. Its recall score of 76.19% highlights its ability to correctly identify actual flood events, ensuring that a significant proportion of true positive cases were detected. Additionally, the F1 score of 80%, a harmonic mean of precision and recall, further validates that the model maintains a solid balance between correctly predicting floods while minimizing false positives and false negatives. This performance suggests that KNN effectively captured underlying flood risk patterns based on climate and soil variables, making it a valuable model for early warning systems and disaster management planning. Alongside the metrics, the confusuion matrix shows how the model classified the flash floods instances.

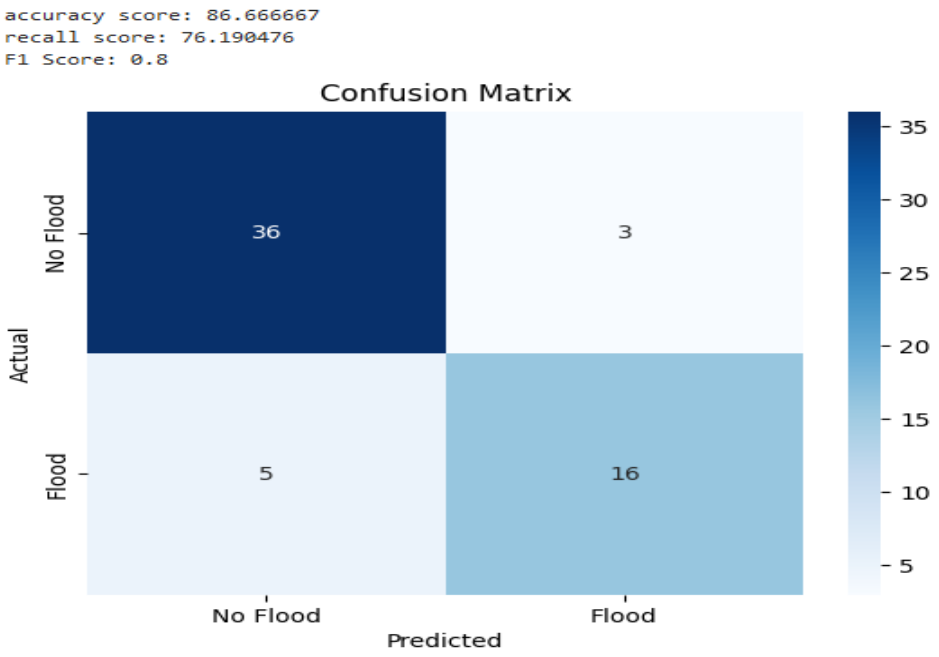


Figure 5.12:KNN Prediction Metrics

Logistic Regression also exhibited strong predictive capabilities achieving an accuracy of 90% as demonstrated in Figure 5.13, indicating that the model made correct classifications in most cases. Its recall score of 71.43% demonstrated its effectiveness in identifying actual flood occurrences, ensuring that a significant portion of true flood events were correctly detected. Additionally, the F1 score of 0.83, which balances precision and recall, confirmed the model's reliability in making accurate predictions while minimizing false positives and false negatives. This performance suggests that Logistic Regression is a robust and interpretable choice for flood prediction, making it suitable for applications where transparency in decision-making is essential.

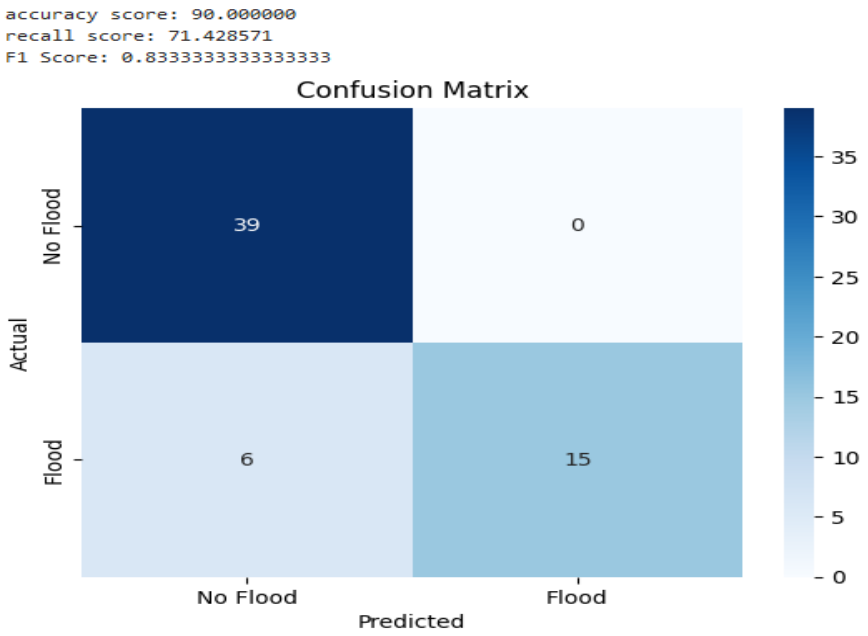


Figure 5.13: LR Prediction Metrics

Additionally, the Artificial Neural Network (ANN) model initially showed modest performance, with accuracy fluctuating between 63.57% and 80%. However, as the model was trained over 100 epochs, it demonstrated significant improvement, achieving an accuracy of 87% at epoch 45, with a validation accuracy of 95%. This improvement over time highlighted the ANN's capacity to learn effectively from the data and enhance its predictions. Therefore, Random Forest outperformed all other models, demonstrating the highest accuracy and recall. Logistic Regression, SVM, KNN, and the ANN models offered competitive results. However, given its high accuracy, recall, and overall balanced performance, Random Forest was identified as the most effective model for predicting flash floods in Nairobi County.

Chapter 6: Discussions

6.1 Interpretation of the Findings Related to Research Questions

The findings revealed that rainfall, soil moisture content, erosion degree, and river discharge were the most significant factors influencing flood occurrence, as indicated by their strong positive correlations with the target variable. Rainfall exhibited the highest correlation with flood occurrence, followed by soil moisture content, then river discharge, and finally by erosion degree which showed lower correlation. Regarding model performance, the Random Forest and Support Vector Machine models achieved the highest accuracy at 93%, demonstrating their effectiveness in predicting flood events. Logistic Regression also performed well with 90% accuracy, while K-Nearest Neighbors and Artificial Neural Networks achieved lower accuracies of 86%. These results highlighted the importance of selecting robust models and key predictors for accurate flood prediction.

6.1.1 Real World Use Case and Testing

The models' real world use cases were tested in accordance with the system designed where the findings in terms of factors that affects flash flood occurrence were used to make prediction in situations where such data was available. For the soil moisture content, the percentages of soil moisture were set to range from 21.9% to 55.4% since these are the lowest and the highest values of soil moisture which have ever been recorded in Nairobi (Ngugi et al., 2019). Rainfall amount used for testing the model was between 0.5 millimetres and 750 millimetres since the least rainfall ever recorded in the world was 0.5 millimetres and the highest for Nairobi is 750 millimetres (Kilavi et al. 2018; University of Cape Town, 2017). River discharge volumes used were between 0.20 and 2.03 cubic meters per second. This was because these are the values ever recorded for Nairobi River according to Masibo (1990). Erosion degree which quantifies how much soil has been lost due to erosion (Kong et al., 2022) was the last feature used. The variation of erosion degree used were slight of above 75%, moderate of between 25% to 75% and severe of less than 25% soil retention on the surface. The erosion degree classifications were figures suggested by the Indiana Soils; Evaluation and Conservation Online Manual n.d and Zenebe et al. (2022). In the real-world application, the models were found to be able to make accurate predictions and therefore suitable for application for Nairobi County.

6.2 Discussion of the Findings in Comparison with Literature

The current study's findings on flash flood prediction in Nairobi County revealed that rainfall, soil moisture content, erosion degree, and river discharge were key predictors, with Random Forest and Support Vector Machine models achieving the highest accuracy (93%). When compared to Wu et al. (2023) study, which used a probabilistic approach integrating Flash Flood Guidance and Kernel Density Estimation for flood risk classification, this study focused on binary classification through machine learning models trained on historical data. Similarly, Lin et al. (2022) employed traditional hydrological modeling (HEC-HMS) for continuous forecasting metrics, while the current study adopted a data-driven approach emphasizing classification accuracy.

The study closely aligns with Al-Rawas et al. (2024) study, which highlighted the effectiveness of machine learning models like Random Forest and SVM in flood prediction and emphasized the integration of AI with IoT and cloud technologies for real-time warning systems. Comparatively, Wang et al. (2024) demonstrated the potential of deep learning, with Temporal Convolutional Networks outperforming CNNs in capturing spatiotemporal flood dynamics. Incorporating such models could enhance the current study's predictive capabilities. Kilavi et al. (2018) emphasized the importance of sub-seasonal forecasts and climatic drivers of floods in Kenya, suggesting that integrating meteorological and socio-economic data could further improve model performance and urban flood risk management.

6.2.1 Key Factors Influencing Flood Occurrence in Nairobi County

The study found that rainfall, soil moisture content, river discharge, and erosion degree were the most significant factors influencing flood occurrence in Nairobi County. Among these, rainfall showed the highest positive correlation with flood events, affirming its central role in triggering flash floods. This aligns with expectations in hydrological modeling, as intense or prolonged rainfall overwhelms drainage systems, especially in urban areas. Soil moisture content, which followed rainfall in significance, indicates that saturated soils have a reduced capacity to absorb additional precipitation, thereby increasing surface runoff. This result emphasizes the compound effect of pre-existing soil conditions on flood risk, a factor sometimes underrepresented in conventional flood models.

River discharge was also positively correlated, reinforcing the contribution of catchment-level hydrological behaviour to localized flooding. Erosion degree, while positively correlated, showed a weaker relationship. This suggests that although erosion may exacerbate sedimentation and reduce channel capacity, its direct impact on immediate flood occurrence is

less significant compared to dynamic hydrometeorological variables. In contrast to some previous studies, such as Kilavi et al. (2018), which emphasized sub-seasonal rainfall variability and topography, this study provides a more data-driven hierarchy of flood determinants, with a clear dominance of rainfall and soil moisture content. These findings suggest that real-time monitoring of these two variables could significantly improve flood forecasting systems in urban Kenyan settings.

6.2.2 Machine Learning Models in Predicting Floods

Among the five machine learning models tested, Random Forest (RF) and Support Vector Machine (SVM) achieved the highest accuracy, both at 93%, followed by Logistic Regression (90%), while K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN) performed less effectively with accuracies of 86%. These results highlight the importance of model selection in flood prediction tasks, particularly in data-limited or highly variable environments like Nairobi County. The strong performance of RF and SVM can be attributed to their ability to handle complex, non-linear relationships and interactions between variables—a common feature in hydrological data. RF also provides feature importance rankings, helping identify the most influential flood predictors, which aligns well with the study's first objective.

In contrast, the relatively lower performance of ANN and KNN may be due to the small or medium-sized dataset, where neural networks are prone to overfitting or require more training data to reach optimal generalization. These results differ from studies like Wang et al. (2024), where ANN models outperformed others, likely due to larger datasets or temporal modeling capacity. In this study, simpler and more interpretable models like RF and SVM were more suited to the dataset and objectives. These findings underscore that accuracy alone should not guide model selection—interpretability, scalability, and data availability are also critical considerations in practical flood forecasting systems.

6.3 Implications Related to the Research Questions

The findings of this study provide critical insights for improving flood prediction and risk management strategies in Nairobi County. By identifying rainfall, soil moisture content, erosion degree, and river discharge as the most influential variables in flash flood occurrence—with rainfall showing the strongest correlation—the study offers actionable guidance for prioritizing monitoring and early warning efforts. These findings directly respond to the study's core question on which environmental factors are most predictive of flash flood events. The

superior performance of machine learning models, particularly Random Forest and Support Vector Machine (both achieving 93% accuracy), confirms their suitability for real-time flood detection and alerts, thus enabling more effective early warning systems. This can support emergency response teams in issuing timely alerts, reducing disaster response time, and mitigating the loss of life and property.

In terms of policy and planning, the results underscore the need for an integrated flood management approach that combines predictive modeling with urban planning, infrastructure development, and public awareness. Decision-makers can leverage these findings to formulate zoning regulations that restrict construction in flood-prone areas, develop urban infrastructure such as retention basins and permeable pavements to manage runoff, and reinforce river embankments to contain overflow. Resource allocation can also be optimized by focusing investments on monitoring systems for rainfall, river discharge, and soil moisture in vulnerable zones. Furthermore, community-level education and outreach programs can be tailored around these key risk factors, empowering residents with knowledge on how to respond to flood warnings and implement localized mitigation strategies. Overall, the study not only provides a technical foundation for enhancing flood prediction but also offers a roadmap for coordinated, data-driven flood risk reduction in Nairobi County.

6.4 Contribution of the Study

6.4.1 Contribution to Academic Literature

This study makes a significant contribution to academic literature by advancing the application of machine learning techniques in flash flood prediction. By integrating climate variables such as temperature, rainfall, and wind speed with soil and terrain characteristics, this research highlights the effectiveness of data-driven models in identifying flood-prone areas. The study compares multiple classification algorithms, including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), providing valuable insights into their predictive capabilities and limitations. Additionally, the research enhances the existing body of knowledge on feature selection techniques, demonstrating the impact of various environmental factors on flood risk. By utilizing Python 3.9 and its data analysis and machine learning libraries for model development, this study contributes methodological insights into implementing machine learning models in geospatial and environmental sciences, bridging the gap between computational intelligence and disaster risk management. Future researchers can build on these findings to improve

predictive accuracy, refine feature selection techniques, and explore hybrid modeling approaches for more robust flood forecasting.

6.4.2 Contribution to Risk Management

From a flood risk management perspective, this study provides a data-driven framework that enhances early warning systems and disaster preparedness. By accurately predicting flood occurrences, local authorities, policymakers, and disaster response teams can make informed decisions regarding evacuation plans, infrastructure development, and resource allocation in flood-prone areas. The insights from this model can help in prioritizing regions that require urgent interventions, such as improved drainage systems or reforestation efforts to mitigate flood risks. Furthermore, integrating the model into real-time monitoring systems can allow for proactive disaster response, minimizing economic losses and safeguarding lives. By demonstrating how machine learning can be leveraged for risk assessment, this study promotes the adoption of advanced predictive analytics in disaster management, offering a scalable and adaptable solution for urban planning and climate resilience strategies in Nairobi County and beyond.

6.5 Limitations of the Study

Reflecting on the research process reveals both challenges and successes. Data limitations, especially for environmental and infrastructural variables, posed a constraint. Data on river management, urbanization, deforestation, encroachment, and drainage systems was not readily available. However, the study achieved key successes, such as effectively evaluating machine learning models and identifying Random Forest as the most reliable for flood prediction. Comprehensive insights from the correlation analysis enhanced understanding of variable importance, and the findings have immediate practical applications in disaster management and urban planning. Addressing these challenges in future research could involve collecting more granular, real-time data, fostering interdisciplinary collaboration with urban planners, hydrologists, and policymakers, and developing user-friendly models.

Chapter 7: Conclusions and Recommendations

7.1 Conclusions

The study concluded that rainfall, soil moisture, river discharge, and erosion degree are the most significant factors influencing flash flood occurrences in Nairobi County, while land use type, humidity, and temperature had a moderate impact, and variables such as elevation, surface stoniness, and slope position showed minimal influence. Using these identified factors, several predictive models were developed and evaluated to determine their effectiveness in forecasting flash flood events. Among the tested models, Random Forest emerged as the most accurate and reliable, achieving a high accuracy of 93.33%, recall of 90.47%, ROC score of 91.58%, and an F1 score of 0.90. Although Logistic Regression and Support Vector Machine (SVM) also performed well, with accuracies of 90% and 93.67% respectively, Random Forest offered the most balanced performance across all evaluation metrics. K-Nearest Neighbors (KNN) provided a dependable yet less optimal outcome, while the Artificial Neural Network (ANN) model showed promising improvements after additional training, reaching 86.43% accuracy after 100 epochs. These findings are important because they demonstrate the feasibility of using machine learning models particularly Random Forest to accurately predict flash floods, which can play a crucial role in early warning systems, disaster preparedness, and flood risk mitigation. By understanding the key contributing factors and employing the most effective predictive tools, stakeholders can take informed, proactive measures to minimize the impact of flash floods, ultimately protecting lives, infrastructure, and resources in flood-prone regions like Nairobi County.

7.2 Recommendations

Based on the study's findings, it was recommended that local authorities and disaster management agencies integrate the Random Forest predictive model into their early warning systems for flash floods in Nairobi County. This model's high accuracy and balanced performance make it a reliable tool for forecasting flood risks, enabling timely alerts and effective emergency response planning to safeguard lives and property.

Secondly, continuous monitoring and real-time data collection of key variables—particularly rainfall, soil moisture, river discharge, and erosion degree should be prioritized. Establishing a robust network of weather stations, soil sensors, and river gauging tools will ensure that the predictive model receives up-to-date information, enhancing its accuracy and responsiveness in real-world scenarios.

Lastly, community awareness and preparedness programs should be strengthened to complement technological interventions. Educating residents in flood-prone areas about the signs of flash floods, evacuation procedures, and the availability of early warnings will empower them to take swift action during emergencies, thereby reducing the human and economic toll of flood events.

7.3 Future Research Work

Future research could build on the findings of this study by addressing several areas. Additional variables, such as urban planning metrics, vegetation cover, and climate change projections, could provide a more comprehensive understanding of flash flood drivers. Longitudinal analyses examining flood occurrence trends, particularly concerning urbanization and climate variability, would enhance our knowledge of temporal patterns. Expanding the study to include regional comparisons across multiple counties or regions could reveal spatial variability in flash flood predictors and improve the generalizability of the models. Additionally, future researchers can explore hybrid modelling approaches for more robust flood forecasting.

References

2018 Revision of World Urbanization Prospects. (2018). United Nations. Retrieved December 1, 2024, from https://www.un.org

Abdouli, N., Hussein, N., Ghebreyesus, N., & Sharif, N. (2019). Coastal runoff in the United Arab Emirates—The hazard and opportunity. *Sustainability*, *11*(19), 5406. https://doi.org/10.3390/su11195406

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics, 8. https://doi.org/10.3389/fninf.2014.00014

Achawakorn, K., Raksa, K., & Kongkalai, N. (2014). Flash flood warning system using SCADA system: Laboratory level. https://doi.org/10.1109/ieecon.2014.6925908

Adan, H. (2024, May 22). UN says 1.6 million Eastern Africans so far affected by heavy rains, floods. Eastleigh Voice. https://eastleighvoice.co.ke/regional/43731/un-says-1-6-million-eastern-africans-so-far-affected-by-heavy-rains-floods.

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61-75.

Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a Web-Based Prediction System for Students' Academic Performance. Data 2022, 7, 21.

Al-Rawas, G., Nikoo, M. R., Al-Wardy, M., & Etri, T. (2024). A critical review of emerging technologies for flash flood prediction: examining artificial intelligence, machine learning, internet of things, cloud computing, and robotics techniques. Water, 16(14). https://doi.org/10.3390/w16142069

Alsaqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile Software Development: Methodologies and Trends. International Journal of Interactive Mobile Technologies (iJIM), 14(11), 246. https://doi.org/10.3991/ijim.v14i11.13269

Ammann, P., & Offutt, J. (2016). Introduction to software testing. https://doi.org/10.1017/9781316771273

Ankita Bhattacharya. (2022, April 20). What is Hybrid Machine Learning and How to Use it? Analytics Insight. https://www.analyticsinsight.net/latest-news/what-is-hybrid-machine-learning-and-how-to-use-it

Ayugi, B., Tan, G., Niu, R., Dong, Z., Ojara, M., Mumo, L., Babaousmail, H., & Ongoma, V. (2020). Evaluation of Meteorological Drought and Flood Scenarios over Kenya, East Africa. Atmosphere, 11(3), 307. https://doi.org/10.3390/atmos11030307

Bates, P., & De Roo, A. (2000). A simple raster-based model for flood inundation simulation. Journal of Hydrology, 236(1–2), 54–77. https://doi.org/10.1016/s0022-1694(00)00278-x

BBC News. (2024, February 8). What is climate change? A really simple guide. https://www.bbc.com/news/science-environment-24021772

Bibi, T. S., Reddythta, D., & Kebebew, A. S. (2023). Assessment of the drainage systems performance in response to future scenarios and flood mitigation measures using stormwater management model. City and Environment Interactions, 19, 100111. https://doi.org/10.1016/j.cacint.2023.100111

Bithari, T. B., Thapa, S., & Hari, K. C. (2020). Predicting academic performance of engineering students using ensemble method. *Technical Journal*, *2*(1), 89-98.

Blockeel, H., Devos, L., Frenay, B., Nanfack, G., & Nijessen, S. (2023). Decision Trees: From efficient prediction to responsible AI. Horizons in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1124553

Brakenridge, G. R., Syvitski, J. P. M., Overeem, I., Higgins, S. A., Kettner, A. J., Stewart-Moore, J. A., & Westerhoff, R. (2012). Global mapping of storm surges and the assessment of coastal vulnerability. Natural Hazards, 66(3), 1295–1312. https://doi.org/10.1007/s11069-012-0317-z

Breskuvienė, D., & Dzemyda, G. (2023). Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. *International Journal of Computers Communications & Control*, *18*(3).

Brownlee, J. (2020, August 26). Train-Test split for evaluating machine learning algorithms. Machine Learning Mastery. Retrieved September 14, 2024, from https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Brownlee, J. (2020a, August 17). *Ordinal and One-Hot encodings for categorical data*. Machine Learning Mastery. Retrieved October 10, 2024, from https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/

Brownlee, J. (2020b, August 28). *How to scale data with outliers for Machine Learning*. Machine Learning Mastery. Retrieved October 20, 2024, from https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/

Carter, J., & Singam, C. (2024, November 24). *System Architecture Design Definition*. SEBoK Wiki. https://sebokwiki.org/wiki/System_Architecture_Design_Definition

Cheng, Y., Huang, X., Li, C., & Shen, Z. (2016). Field and numerical investigation of soil–atmosphere interaction at Nairobi, Kenya. European Journal of Environmental and Civil Engineering, 21(11), 1326–1340. https://doi.org/10.1080/19648189.2016.1169224

Chouhbi, K. (2020, July 8). Preprocessing Data: feature scaling - towards data science. *Medium*. https://towardsdatascience.com/preprocessing-data-feature-scaling-cc28c508e8af

Clarke, B., Otto, F., Stuart-Smith, R., & Harrington, L. (2022). Extreme weather impacts of climate change: an attribution perspective. Environmental Research Climate, 1(1), 012001. https://doi.org/10.1088/2752-5295/ac6e7d

Dahri, N., & Abida, H. (2020). Causes and impacts of flash floods: case of Gabes City, southern Tunisia. Arabian Journal of Geosciences, 13, 176. https://doi.org/10.1007/s12517-020-5149-7

Dash, R. K., Nguyen, T. N., Cengiz, K., & Sharma, A. (2023). Fine-tuned support vector regression model for stock predictions. *Neural Computing and Applications*, *35*(32), 23295-23309.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. BMC Medical Genomics, 4(1). https://doi.org/10.1186/1755-8794-4-31

Doerr, S., Shakesby, R., & Walsh, R. (2000). Soil water repellency: its causes, characteristics and hydro-geomorphological significance. *Earth-Science Reviews*, 51(1–4), 33–65. https://doi.org/10.1016/s0012-8252(00)00011-8

Fawcett, A. (2024, May 29). *Data science in 5 minutes: What is one hot encoding?* Educative. Retrieved October 10, 2024, from https://www.educative.io/blog/one-hot-encoding

Ferreira, C. S. S., Potočki, K., Kapović-Solomun, M., & Kalantari, Z. (2021). Nature-Based solutions for flood mitigation and resilience in urban areas. In The handbook of environmental chemistry (pp. 59–78). https://doi.org/10.1007/698_2021_758

Gauhar, N., Das, S., & Moury, K. S. (2021). Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm. 2021 2nd International Conference on Robotics, Electrical and

Signal Processing Techniques (ICREST). https://doi.org/10.1109/icrest51555.2021.9331199

GeeksforGeeks. (2024, January 10). *System analysis | system design*. Retrieved January 21, 2025, from https://www.geeksforgeeks.org/system-analysis-system-design/

GeeksforGeeks. (2025, January 8). *Functional vs. non-functional requirements*. Retrieved January 21, 2025, from https://www.geeksforgeeks.org/functional-vs-non-functional-requirements/

GeeksforGeeks. (2025, January 3). *Sequence Diagrams – Unified Modeling Language (UML)*. GeeksforGeeks. https://www.geeksforgeeks.org/unified-modeling-language-uml-sequence-diagrams/

Gray, S. (2021, December 15). Agile Software Development Life Cycle - Serena Gray – Medium. Medium. https://serenagray2451.medium.com/agile-software-development-life-cycle-b3ed0f0f7212

Grignon, K. M. (2024). How impunity and corruption fuel disasters and claim Kenyan lives. The Standard. https://www.standardmedia.co.ke/health/opinion/article/2001502698/how-impunity-and-corruption-fuel-disasters-and-claim-kenyan-lives

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan Kaufmann.

Hoc, H. T., Silhavy, R., Prokopová, Z., & Silhavy, P. (2023). Comparing stacking ensemble and deep learning for software project effort estimation. *Research Gate*. https://doi.org/10.1109/ACCESS.2023.3286372

Hope, T. M. (2020). Linear regression. In *Machine learning* (pp. 67-81). Academic Press.

Igini, M. (2023, September 20). Climate Change Made Libya Floods Up to 50 Times More Likely, Study Finds

Indiana Soils:Evaluation and Conservation Online Manual. (n.d.). https://www.agry.purdue.edu/soils_judging/new_manual/Ch2-erosion.html

Imambi, S., Prakash, K. B., & Kanagachidambaresan, G. R. (2021). PyTorch. In EAI/Springer Innovations in Communication and Computing (pp. 87–104). https://doi.org/10.1007/978-3-030-57077-4_10 .

Implementation of Nature-Based Solutions for climate resilient and flood risk Management in Pakistan | Department of Economic and Social Affairs. (2023, March 6). https://sdgs.un.org/partnerships/implementation-nature-based-solutions-climate-resilient-and-flood-risk

management#:~:text=Wetlands%20act%20as%20natural%20buffers,biodiversity%2C%20and%20providing%20recreational%20opportunities.

Intharasombat, O., & Khoenkaw, P. (2015). A low-cost flash flood monitoring system. https://doi.org/10.1109/iciteed.2015.7408993

Jardosh, P., Kanvinde, A., Dixit, A., & Dholay, S. (2020). Detection of flood prone areas by flood mapping of SAR imagery. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). https://doi.org/10.1109/icssit48917.2020.9214089

Juma, B., Olang, L. O., Hassan, M., Chasia, S., Bukachi, V., Shiundu, P., & Mulligan, J. (2021). Analysis of rainfall extremes in the Ngong River Basin of Kenya: Towards integrated urban flood risk management. Physics and Chemistry of the Earth Parts a/B/C, 124, 102929. https://doi.org/10.1016/j.pce.2020.102929

Kaburu, M., Koech, M., & Manguriu, D. (2019). ANTHROPOGENIC FACTORS THAT CAUSE FLOODS IN MUKURU SLUMS, NAIROBI CITY COUNTY, KENYA. European International Journal of Science and Technology, 8(7). https://doi.org/10.3390/atmos11030307

Kamiri, J., & Mariga, G. (2021). Research Methods in Machine Learning: A content analysis. International Journal of Computer and Information Technology (2279-0764), 10(2). https://doi.org/10.24203/ijcit.v10i2.79

Kanade, V. (2022, May 30). Decision Tree Algorithms, Template, Best Practices – Spiceworks Inc. Spiceworks Inc. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-decision-tree/

Kenya Red Cross. (2024). Floods operations, 2024. https://www.redcross.or.ke/floods/

Kilavi, M., MacLeod, D., Ambani, M., Robbins, J., Dankers, R., Graham, R., Titley, H., Salih, A., & Todd, M. (2018). Extreme Rainfall and Flooding over Central Kenya Including Nairobi City during the Long-Rains Season 2018: Causes, Predictability, and Potential for Early Warning and Actions. Atmosphere, 9(12), 472. https://doi.org/10.3390/atmos9120472

Kim, N., Kim, S., Sung, K. K., & Jae, L. K. (2008). Evaluation of Effects on SWAT

Kong, H., Wu, D., & Yang, L. (2022). Quantification of soil erosion in small watersheds on the Loess Plateau based on a modified soil loss model. *Water Science & Technology Water Supply*, *22*(7), 6308–6320. https://doi.org/10.2166/ws.2022.256

Kumar, R. (1999). Research Methodology: A Step-by-Step Guide for Beginners. http://library.pps.uny.ac.id/opac/index.php?p=show_detail&id=8748&keywords=

Kumar, S. (2023, January 19). *The concept behind "Mean Target encoding" in AI & ML*. Retrieved October 10, 2024, from https://hackernoon.com/the-concept-behind-mean-target-encoding-in-ai-and-ml

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., Mechler, R., Bouwer, L. M., Arnell, N., Mach, K., Muir-Wood, R., Brakenridge, G. R., Kron, W., Benito, G., Honda, Y., Takahashi, K., & Sherstyukov, B. (2013). Flood risk and climate change: global and regional perspectives. Hydrological Sciences Journal, 59(1), 1–28. https://doi.org/10.1080/02626667.2013.857411

Lai, O. (2023, September 30). What are the main causes and effects of floods around the world? Earth.org. https://earth.org/what-are-the-main-causes-and-effects-of-floods/

Laudan, J., Zöller, G., & Thieken, A. H. (2020). Flash floods versus river floods – a comparison of psychological impacts and implications for precautionary behaviour. Natural Hazards and Earth System Sciences, 20(4), 999–1023. https://doi.org/10.5194/nhess-20-999-2020

Lin, Q., Lin, B., Zhang, D., & Wu, J. (2022). Web-based prototype system for flood simulation and forecasting based on the HEC-HMS model. Environmental Modelling & Software, 158, 105541.

Li, H., Liao, W., & Le, M. (2021). Integrated Risk Analysis Rely on Multi-data for Flash Flood in China Considering the Sensitivity of Disaster-prone Environment. https://doi.org/10.1109/icbar55169.2021.00024

Masibo, M. (1990). *A study of the hydrogeology, solid waste disposal and river water pollution in the Nairobi area* [MA thesis]. University of Nairobi.

MaxAbsScaler. (n.d.). Scikit-learn. Retrieved October 29, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html

Middleton, F. (2019, August 8). The 4 Types of Reliability in Research | Definitions & Examples. Scribbr. Retrieved September 11, 2024, from https://www.scribbr.com/methodology/types-of-reliability/

MinMaxScaler. (n.d.). Scikit-learn. Retrieved October 10, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Mishra, V. (2024, August 29). *Floods, landslides wreak havoc across South Asia*. United Nations. https://news.un.org/en/story/2024/08/1153726

Mabonga, M., & Amondi, A. (2024, December 13). *Preventing Disasters: How Kenya's Meteorological Department tackles floods*. Tuko.co.ke - Kenya News.

https://www.tuko.co.ke/editorial/analysis/570279-preventing-disasters-how-kenyas-meteorological-department-tackles-floods/

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature review. Water, 10(11), 1536. https://doi.org/10.3390/w10111536

Mugenda, O. M., & Mugenda, A. G. (2003). Research methods, qualitative and quantitative approaches. Nairobi: African Centre for Technology Studies.

Mwaisaka, M. (2023, November 29). Flood risk occurrence in Nairobi. ArcGIS StoryMaps. https://storymaps.arcgis.com/stories/1a7e3db01a44418c865d781ebfdfee8b

Mwangangi, S. (2023, October 2). The safest foundation to build on black cotton soil. https://www.linkedin.com/pulse/safest-foundation-build-black-cotton-soil-stanley-mwangangi/

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean journal of anesthesiology*, *75*(1), 25-36.

NASA Langley Research Centre. (2024). POWER Data Access Viewer. NASA Langley Research Centre. https://power.larc.nasa.gov. Accessed September 20, 2024.

National Severe Storms Laboratory. (n.d.). *Flood basics*. NOAA. https://www.nssl.noaa.gov/education/svrwx101/floods/types/

Ngugi, H. N., Shitote, S., Ambassah, N., Okumu, V., & Thuo, J. (2019). Influence of variation in moisture content to soil bearing capacity in Nairobi area and its environs. *American Journal of Engineering and Technology Management*, *4*(6), 97. https://doi.org/10.11648/j.ajetm.20190406.14

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 52(1), 77–124. https://doi.org/10.1007/s10462-018-09679-z

Nhita, F., Adiwijaya, Annisa, S., & Kinasih, S. (2015). Comparative study of Grammatical evolution and Adaptive Neuro-fuzzy Inference System on rainfall forecasting in Bandung. IEEE. https://doi.org/10.1109/icoict.2015.7231388

Ninja, N. (2024, April 1). *Frequency encoding: counting categories for representation*. Let's Data Science. https://letsdatascience.com/frequency-encoding/

Njogu, H. W. (2021). Effects of floods on infrastructure users in Kenya. Journal of Flood Risk Management, 14(4). https://doi.org/10.1111/jfr3.12746.

NOAA, no date. **Flash Flooding Definition**. National Oceanic and Atmospheric Administration (NOAA), National Weather Service. Accessed 16 October 2024.

OCHA. (2024, May 30). Eastern Africa: Heavy rains and flooding Flash Update #4 (30 May 2024). United Nations Office for the Coordination of Humanitarian Affairs. https://www.unocha.org/publications/report/kenya/eastern-africa-heavy-rains-and-flooding-flash-update-4-30-may 2024#:~:text=Heavy%20rains%20and%20flash%20floods,displaced%2C%20as%20of%2030%20May

Okaka, F. O., & Odhiambo, B. (2018). Relationship between flooding and out break of infectious diseases in Kenya: A review of the literature. Journal of Environmental and Public Health, 2018. https://doi.org/10.1155/2018/5452938

Pal, S. (2023, September 13). What is StandardScaler - How & Why We Use. *GeekPython - Python Programming Tutorials*. Retrieved October 25, 2024, from https://geekpython.in/how-to-use-sklearn-standardscaler#google_vignette

Pham, B. T., Jaafari, A., Van Phong, T., Yen, H. P. H., Tuyen, T. T., Van Luong, V., Nguyen, H. D., Van Le, H., & Foong, L. K. (2021). Improved flood susceptibility mapping using a best first decision tree integrated with ensemble learning techniques. Geoscience Frontiers, 12(3), 101105. https://doi.org/10.1016/j.gsf.2020.11.003

Ponnamperuma, N., & Rajapakse, L. (Eds.). (2020). Holistic behaviour of urban pond systems for flood risk mitigation- a case study in Metro Colombo area. IEEE. https://doi.org/10.1109/MERCon50084.2020.9185383

Ramiaramanana, F. N., & Teller, J. (2021). Urbanization and Floods in Sub-Saharan Africa: Spatiotemporal Study and Analysis of Vulnerability Factors—Case of Antananarivo Agglomeration (Madagascar). Water, 13(2), 149. https://doi.org/10.3390/w13020149

Requirements.com. (2024, July 12). *What are functional requirements?* Retrieved January 21, 2025, from https://requirements.com/Content/What-is/what-are-functional-requirements

Rukanga, B. B. (2024, April 30). Kenya floods: What a deluge reveals about Nairobi's vulnerability. https://www.bbc.com/news/world-africa-68898731

SciKit learn Tutorial. (n.d.). https://www.tutorialspoint.com/scikit_learn/index.htm

Sharma, A., Sharma, A., Tselykh, A., Bozhenyuk, A., Choudhury, T., Alomar, M. A., & Sánchez-Chero, M. (2023). Artificial intelligence and internet of things oriented sustainable precision farming: Towards modern agriculture. *Open Life Sciences*, *18*(1), 20220713.

Shukla, N. (2021, July 8). *Flooding will hit Asia the hardest: Report*. Earth.org. https://earth.org/climate-change-flooding-will-hit-asia-the-hardest/

Simulated Hydrology and Sediment Behaviors of SWAT Watershed Delineation using SWAT ArcView GIS Extension Patch. Journal of Korean Neuropsychiatric Association, 24(2), 147–155. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001238727

Sky News. (2024, April 24). Devastating flooding in east Africa claims dozens of lives and displaces thousands. https://news.sky.com/story/devastating-flooding-in-east-africa-claims-dozens-of-lives160and-displaces-thousands-13122189?dcmp=snt-sf-twitter

Sonoda, R. (2024, February 8). 3 Key Encoding Techniques for Machine Learning: A Beginner-Friendly Guide with Pros, Cons, and Python Code Examples. *Medium*. https://towardsdatascience.com/3-key-encoding-techniques-for-machine-learning-a-beginner-friendly-guide-aff8a01a7b6a

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101.

Statista. (2024, September 26). Floods - statistics & facts. https://www.statista.com/topics/11261/floods/#topicOverview

Team, C. (2023, May 2). *What is hashing, and how does it work? - Codecademy blog*. Codecademy Blog. https://www.codecademy.com/resources/blog/what-is-hashing/

Thiemig, V., Roo, A., & Gadain, H. (2011). Current status on flood forecasting and early warning in Africa. International Journal of River Basin Management, 9(1). https://doi.org/10.1080/15715124.2011.555082

Todd, M. D. (2022). Sensor data acquisition systems and architectures. In *Elsevier eBooks* (pp. 19–49). https://doi.org/10.1016/b978-0-08-102696-0.00012-9

Tom, R. O., George, K. O., Joanes, A. O., & Haron, A. (2022). Review of flood modelling and models in developing cities and informal settlements: A case of Nairobi city. Journal of Hydrology Regional Studies, 43, 101188. https://doi.org/10.1016/j.ejrh.2022.101188

UNICEF. (2024, May 9). Almost 1 million people in Kenya, Burundi, Tanzania, and Somalia affected as unprecedented heavy rains continue to wreak havoc in Eastern Africa. https://www.unicef.org/press-releases/almost-1-million-people-kenya-burundi-tanzania-and-somalia-affected-unprecedented

University of Cape Town. (2017). Nairobi Climate Profile: Full Technical Version.

Urban development. (2023, April 3). World Bank Group. Retrieved December 5, 2024, from https://www.worldbank.org/en/topic/urbandevelopment/overview

Voice of America. (2024, May 2). Kenya floods death toll rises to 188 as heavy rains persist. https://www.voaafrica.com/a/kenya-floods-death-toll-rises-to-188-as-heavy-rains-persist-/7594986.html

Wainwright, C. M., Finney, D. L., Kilavi, M., Black, E., & Marsham, J. H. (2020). Extreme rainfall in East Africa, October 2019–January 2020 and context under future climate change. Weather, 76(1), 26–31. https://doi.org/10.1002/wea.3824

Wang, X., Goreville, P., & Liu, C. (2023). Flash Floods: Forecasting, monitoring and Mitigation Strategies. Water, 15(9), 1700. https://doi.org/10.3390/w15091700

Wang, X., Xiao, M., Liu, Y., Guo, J., Qin, Y., & Zhang, Y. (2024). A rapid and efficient method for flash flood simulation based on deep learning. Engineering Applications of Computational Fluid Mechanics, 18(1), 2407016.

World Health Organization, Regional Office for Africa. (2024, May 24). Flooding in Ethiopia. https://www.afro.who.int/sites/default/files/2024-05/PHSA%20-Ethiopia%20Floods%20280524%20Final.pdf

World Weather Attribution. (2021, August 23). *Heavy rainfall which led to severe flooding in Western Europe made more likely by climate change.* https://www.worldweatherattribution.org/heavy-rainfall-which-led-to-severe-flooding-in-western-europe-made-more-likely-by-climate-change/

Wu, Z., Bhattacharya, B., Xie, P., & Zevenbergen, C. (2023). Improving flash flood forecasting using a frequentist approach to identify rainfall thresholds for flash flood occurrence. *Stochastic Environmental Research and Risk Assessment*, *37*(1), 429-440.

Xinhua. (2024, May 31). Flooding kills 528 in East Africa amid heavy rains. CGTN Africa. https://africa.cgtn.com/flooding-kills-528-in-east-africa-amid-heavy-rains/

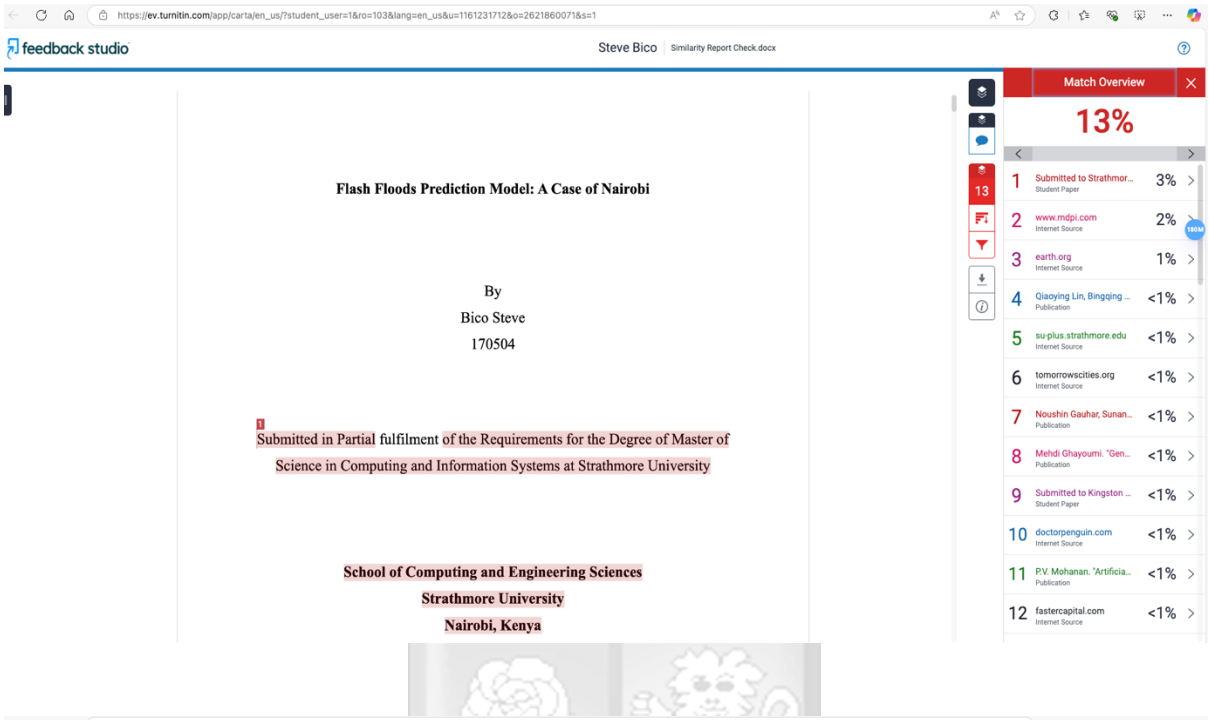
Zafar, Z. (2024). Assessment of urbanization impacts on vegetation cover in major cities of Pakistan: evidence from remotely sensed data. Geo Journal, 89(4). https://doi.org/10.1007/s10708-024-11189-1

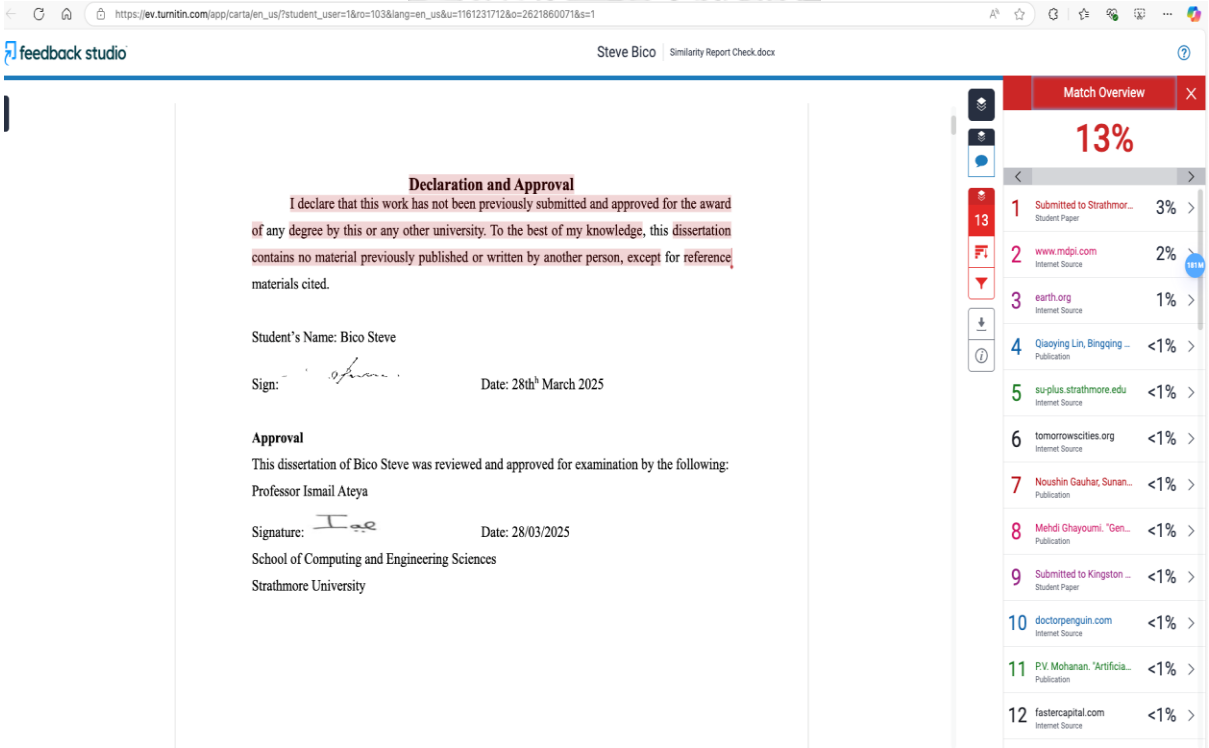
Zenebe, A., Gidey, E., & Hishe, S. (2022). An Integrated Approach of Soil–Erosion Modeling for soil and Water conservation planning in a degraded Semi–Arid Environment of Tigray Region, Northern Ethiopian Highlands. *Modeling Earth Systems and Environment*, *9*(2), 1741–1757. https://doi.org/10.1007/s40808-022-01578-1

Zurich Insurance Group. (2024, April 25). Three common types of floods explained. Natural hazards. https://www.zurich.com/knowledge/topics/flood-and-water-damage/three-common-types-of-flood

Appendices

Appendix A: Similarity Report





Appendix B: Strathmore University Institutional Ethics Review Clearance Certificate

Strathmore UNIVERSITY

26th November 2024

Mr Oloo Bico,
steve.bico@strathmore.edu

Dear Mr Oloo,

RE: Flash Floods Prediction Model: A Case of Nairobi

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2445/24.** The approval period is from **26th November 2024 to 25th November 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C: National Commission for Science, Technology, and Innovation (NACOSTI) Certificate

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **334735**

Date of Issue: **30/December/2024**

RESEARCH LICENSE

This is to Certify that Mr.. Bico Steve of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: Flash Floods Prediction Model: A Case of Nairobi for the period ending : 30/December/2025.

License No: **NACOSTI/P/24/414698**

334735

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 (Rev. 2014)
Legal Notice No. 108: The Science, Technology and Innovation (Research Licensing) Regulations, 2014

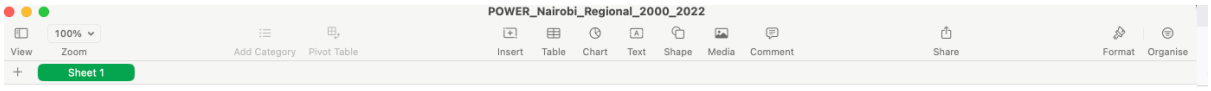
The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was the established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
    i. Endanger national security
    ii. Adversely affect the lives of Kenyans
    iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
    iv. Result in exploitation of intellectual property rights of communities in Kenya
    v. Adversely affect the environment
    vi. Adversely affect the rights of communities
    vii. Endanger public safety and national cohesion
    viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. The license any rights thereunder are non-transferable
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and
Innovation(NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke

Appendix D: NASA Power Nairobi Region Flood Prediction Dataset



POWER_Nairobi_Regional_2000_2022

Slope_Position	Surface_stoniness	Area_affected	Erosion_degree	Sensitivity_to_capping	Wind_Speed_kmh	Temperature	Soil_Moisture_%	Humidity_%	Rainfall_mm	River_Discharge_m3s	Land_Use_Type	Elevation_m	Flood_Occurrence
H	N	1	S	N	28.59979391	25	22.95930502	80.36217965	56.18101783	196.8177601	Forest	1600	0
D	N	4	S	N	40.27161646	24	47.93306632	87.80088383	109.607146	236.7178297	Agriculture	1600	0
H	N	1	V	N	38.00804649	30	79.10620851	55.02807393	142.7990913	427.2736966	Urban	1800	1
L	N	3	V	N	7.694995235	30	61.25574205	77.49244598	130.7987726	270.002193	Agriculture	1600	1
M	N	1	S	N	7.462473487	26	66.45928035	74.30475899	23.40279607	434.8248424	Forest	1600	0
M	N	1	S	N	13.40871839	24	56.11483567	89.96982261	23.39917805	44.06721549	Wetland	1800	0
M	N	2	S	N	18.05373632	24	58.45935952	94.36522363	8.712541825	388.3992186	Agriculture	1900	0
M	N	1	S	N	20.42277907	25	69.44369561	40.72940629	129.9264219	423.7738165	Urban	1900	0
M	N	1	S	N	33.98486084	26	27.47676062	80.44119514	90.16725176	90.90882642	Wetland	1600	0
M	N	1	S	N	2.834021608	23	44.25974746	43.11014795	106.2108867	215.1732662	Agriculture	1600	0
L	N	3	V	N	1.733635211	25	25.48466093	72.9315199	3.087674144	82.74988733	Agriculture	1800	0
M	N	1	V	N	19.59552815	29	79.13676056	57.25796374	145.4864778	353.3020645	Wetland	1800	1
M	M	2	S	N	34.85818461	23	76.08415378	58.40659597	124.8663961	267.673884	Wetland	1900	0
L	N	1	S	W	9.671767148	23	12.7598768	61.17751021	31.8508666	317.6616734	Wetland	1700	0
M	N	3	M	S	32.07522406	25	59.39026208	77.27754694	27.27374508	98.24348348	Urban	1600	0
D	N	5	V	N	12.9914069	28	76.76738222	60.04299794	140.5106765	305.872931	Urban	1700	1
D	N	3	M	N	44.30430544	24	22.64027416	83.96194304	45.63633644	20.74866275	Agriculture	1700	0
M	N	2	S	N	44.78449721	25	49.75616614	64.27164313	78.71346474	161.0337676	Agriculture	1800	0
L	N	2	S	S	14.86436021	26	74.08418083	44.10119202	64.7917528	279.8723307	Forest	1600	0
L	N	1	V	N	11.4996877	22	72.3762185	87.02559055	143.684371	428.967457	Urban	1600	1
A	N	1	M	N	20.5651986	23	58.81941871	57.14549948	91.77793421	333.4637237	Agriculture	1800	0
A	N	1	M	N	12.02657926	24	70.81443052	65.96601278	140.9240791	217.6981617	Wetland	1700	1
A	N	2	V	M	33.61919216	30	74.70773368	81.12663299	143.8216973	476.5600572	Wetland	1800	1
A	N	2	V	M	41.30323423	25	77.97407716	59.94736943	154.9542765	359.6226003	Urban	1900	1
A	N	1	M	N	33.65460655	28	76.09865424	43.39513937	138.4104976	465.0826014	Wetland	1600	1
M	F	2	S	M	41.21752236	23	43.19499517	62.43526282	117.7763942	263.7939866	Forest	1900	0
A	N	1	M	N	19.84960811	26	70.34298557	96.66691496	29.95106732	129.4521374	Agriculture	1800	0
A	N	1	S	N	7.815848535	27	79.1184579	78.50406384	127.1351658	226.4133432	Agriculture	1600	1
A	N	1	S	M	36.89754737	26	32.33703313	80.28874892	88.86218533	363.0424257	Agriculture	1900	0
A	N	1	M	N	18.02372526	26	68.02408319	77.93692922	6.967561908	60.651499	Urban	1900	0
A	N	1	S	N	33.56354126	26	12.59053443	51.93952834	91.13172779	151.3875641	Wetland	1700	0
M	N	1	S	W	13.5321971	26	51.73889149	65.10003185	25.57861855	266.2354539	Forest	1600	0
A	N	5	V	S	4.06148174	26	26.10061861	85.0563879	9.757738948	282.2082939	Urban	1900	0
A	N	2	S	N	49.62908977	28	18.439682	46.08237385	142.3328306	300.2915985	Agriculture	1900	0
A	N	1	S	M	7.810075397	27	15.38672411	56.67116657	144.844805	83.12766517	Forest	1700	0
H	N	1	S	S	49.4210497	26	68.74021431	56.57914651	121.2596022	189.8687024	Agriculture	1900	1
A	N	4	V	S	48.86399663	26	33.79124746	65.92113566	45.69206538	308.5971136	Agriculture	1900	0
A	N	4	V	S	39.6909052	27	70.73367401	98.82212446	114.6508171	484.9094323	Urban	1900	1
H	N	4	S	S	32.9711481	26	14.57494386	44.0501528	102.634954	363.8155688	Agriculture	1700	0
H	N	4	S	S	32.9711481	26	14.57494386	44.0501528	102.634954	363.8155688	Agriculture	1700	0
A	N	5	S	S	28.89035251	27	62.07032365	71.12205966	136.0228741	461.3018785	Forest	1900	1
A	N	4	V	S	43.30507727	27	67.76439047	50.76188092	138.3057352	381.1060292	Agriculture	1900	1
M	N	2	M	M	14.47197608	27	65.35062154	98.24059272	74.27653652	295.8584844	Wetland	1600	0
M	N	1	S	M	23.38406062	28	32.31267521	46.79821028	5.158278167	96.01168457	Agriculture	1600	0
H	N	1	S	S	30.96949847	24	53.81239635	64.21606014	136.3980603	333.3428259	Urban	1700	0
A	C	2	S	N	20.55952379	28	72.01844238	84.27309909	138.8169972	311.689783	Forest	1800	1
L	V	1	M	M	21.3743227	27	53.11042317	82.27326568	99.37834265	301.2345852	Urban	1600	0
A	N	2	V	M	16.51423338	24	76.30716323	65.36371821	146.7566614	244.892934	Wetland	1700	1
A	C	1	S	N	28.21159083	27	11.70805471	60.79145008	78.01020318	264.328302	Agriculture	1900	0
A	N	2	S	S	42.5287264	29	70.90692117	63.85674271	82.0065419	167.2212636	Urban	1900	0
A	N	1	S	S	10.0764189	30	11.48885876	55.85657579	27.72816833	259.6563463	Urban	1700	0
A	N	1	S	N	46.72165194	28	71.22911709	52.31997632	145.4376942	298.7779528	Urban	1900	1
A	N	1	S	M	34.45438269	28	77.02559938	68.98238185	126.2699235	402.5871987	Urban	1900	1
L	N	1	S	S	41.16366063	27	75.7347389	56.11202536	140.9248412	392.8884773	Wetland	1800	1
H	N	3	M	N	27.80953463	26	65.9148265	57.24769989	134.2241026	42.48314494	Wetland	1900	0
M	C	3	M	W	38.97583507	26	79.85538774	79.40536381	89.68499682	218.1230132	Agriculture	1600	0
A	N	2	S	W	0.810019867	28	34.54982708	98.1122399	138.2811353	329.1537198	Agriculture	1900	0
M	C	1	S	N	40.91902124	29	63.70318023	76.21823203	131.2738753	218.8584125	Forest	1700	1
H	M	3	M	W	2.006942954	27	38.13516395	44.61876802	29.39742936	138.377066	Wetland	1800	0
M	N	4	S	S	44.49566882	24	63.59129342	44.53501683	136.7840933	279.0855919	Urban	1900	1
A	M	4	V	W	49.59813736	26	53.92538243	97.08539325	148.7995496	273.5011677	Urban	1700	1
A	M	4	V	M	14.70337311	28	71.15739799	57.83744742	105.3015935	466.4338451	Forest	1600	1
A	C	3	V	M	10.51592781	27	78.88584284	45.52401916	40.70235477	461.3225698	Urban	1900	0
A	A	2	S	S	38.26816819	28	63.77913897	75.94267523	124.3106264	251.18591	Forest	1800	0
M	D	3	V	N	12.65131705	29	69.24367475	77.4189267	153.512999	363.941997	Agriculture	1900	1
M	A	1	S	W	43.27811916	27	39.49499016	78.91028921	42.14017645	368.566207	Wetland	1700	0
M	V	1	S	M	5.142129608	27	61.63076111	56.04412167	81.40441247	18.35647174	Forest	1800	0
A	A	2	S	N	6.297759819	28	26.7144002	40.90664105	21.13863375	237.384503	Urban	1900	0
M	M	3	M	W	48.95757776	29	17.73318792	97.90092217	120.3295471	168.1616971	Forest	1600	0
A	N	5	V	S	33.69194491	27	64.82355103	55.05358274	131.1825966	460.2679426	Agriculture	1800	1
L	F	5	E	S	42.34436122	27	30.10672942	80.56157728	148.0330405	6.001513636	Urban	1600	0
M	N	2	M	M	16.21646887	28	30.74156843	82.39779219	115.8367154	276.4074379	Agriculture	1700	0
A	C	3	M	W	33.82418748	29	26.35254257	76.60044502	29.80735223	370.3317905	Agriculture	1900	0
L	N	2	S	S	29.71209137	29	12.94652327	58.77444365	0.828317569	242.5089938	Forest	1600	0
H	V	1	S	N	30.15750601	27	11.25117543	56.26577523	122.3192143	42.70151508	Agriculture	1700	0
A	M	2	V	M	34.12662298	28	79.14056728	75.86009696	136.0286016	486.230695	Urban	1900	1
A	M	2	M	M	28.76795866	29	39.94411936	91.96573802	109.3510752	259.0052157	Agriculture	1700	0
A	F	1	S	N	21.45311847	29	36.9028653	96.80402374	115.690552	307.0931222	Agriculture	1600	0
A	C	3	M	S	13.79613472	29	57.57530979	46.35434826	11.10669776	118.3048914	Forest	1700	0
A	M	5	V	S	38.42906655	28	25.27777215	49.28971676	53.76985928	241.7490426	Agriculture	1800	0
A	M	1	S	N	11.3136215	29	76.49728288	96.68417553	127.3803589	214.574746	Wetland	1900	1

A	N	2	S	W	34.61771531	29	65.04415101	84.1921137	129.4655139	37.447921	Forest	1700	0
A	N	2	S	W	11.66664113	30	66.25877016	92.97962949	133.494719	453.0801768	Forest	1600	1
L	N	3	V	W	31.2665085	26	69.2306543	52.15795805	149.6347037	418.7362598	Urban	1900	1
M	N	1	S	N	37.35391336	27	71.53828153	75.25515227	9.533752543	119.9099761	Urban	1800	0
H	N	1	S	N	10.93569445	28	76.13124156	82.06837567	46.64734826	97.47913575	Urban	1900	0
H	N	1	S	N	2.997165043	28	42.71810579	80.80671313	48.7774983	252.4557256	Urban	1800	0
H	N	1	S	N	6.54868715	27	52.93879724	64.48910145	109.4409268	384.3735657	Wetland	1900	0
H	N	1	S	N	30.29984836	27	21.69237623	40.9236935	95.6336207	30.97325929	Forest	1700	0
M	C	2	M	N	42.47232031	28	79.38180383	74.97556341	133.0819114	288.4039598	Agriculture	1700	0
D	N	5	S	N	2.249841062	26	26.2170191	55.18609235	70.83223877	59.60522142	Urban	1600	0
M	C	1	S	W	36.70756295	27	75.99122419	67.01525477	17.93913689	18.23052541	Urban	1800	0
M	N	1	S	M	17.0659323	24	55.47526543	97.45486091	136.9867181	326.3719761	Urban	1600	1
D	N	5	V	N	23.9264185	27	52.54157564	63.94211906	114.1177573	417.0820541	Agriculture	1700	0
H	N	1	S	N	46.44297749	27	45.88819577	90.3880964	84.19157964	58.98624181	Urban	1800	0
M	N	1	S	N	16.59857524	22	26.14688682	51.31243634	115.645077	22.57568221	Agriculture	1600	0
L	N	1	S	N	23.2668467	25	22.35696224	80.34762988	74.06933945	218.9499436	Forest	1600	0
M	F	2	V	M	0.684120738	26	65.43403463	98.62041727	128.4099244	421.996736	Forest	1600	1
M	C	3	M	M	4.079936	26	23.05067835	46.11358605	64.13115275	131.2041882	Forest	1600	0
L	N	2	S	W	12.92957956	24	64.57091315	40.49919653	3.812869012	210.8563088	Wetland	1600	0
M	N	3	V	W	1.392612041	25	64.50876814	66.01496756	116.183714	319.9612663	Agriculture	1800	1
M	M	1	S	W	31.5687663	27	14.04898736	45.55752898	4.714377853	224.5392571	Urban	1600	0
M	D	1	S	N	21.31991784	25	77.83718411	84.90306098	95.46156169	288.8065403	Urban	1800	0
M	D	2	V	N	27.40726014	27	71.86501195	94.87291983	117.1533972	285.4535269	Urban	1600	1
M	V	1	S	N	8.732359295	26	74.94265982	66.04125836	76.28560367	166.2133461	Urban	1700	0
M	D	2	M	S	14.79660147	23	79.64354759	55.52269853	136.1349711	157.7665903	Wetland	1700	0
M	A	1	S	W	33.19122374	23	22.17266745	66.06419914	37.39383437	53.31180947	Urban	1600	0
M	F	2	M	M	48.26513618	26	37.73694132	83.40686907	61.55743846	183.3991603	Wetland	1900	0
H	C	2	S	M	2.52083868	26	63.0766933	40.54324969	113.3326708	49.74733444	Agriculture	1700	0
L	N	1	S	S	44.51921604	26	58.72144326	75.36723937	134.3197248	383.2979724	Forest	1800	1
A	M	1	S	S	28.84406007	26	20.77271344	76.79750109	11.54698647	483.1371489	Urban	1600	0
M	C	3	V	M	28.19597834	26	67.10831875	78.26129481	143.4627179	485.0372283	Forest	1800	1
H	N	3	S	M	25.01392676	28	55.71084003	54.5213248	124.1831931	432.5355497	Agriculture	1900	1
L	N	2	S	M	3.469704295	25	25.66723304	82.8431634	139.4546479	300.3960747	Agriculture	1600	0
H	C	1	S	N	4.493842587	25	47.5882096	45.48348918	121.2180569	350.4328486	Urban	1700	0
M	N	4	V	M	30.04560028	25	51.50579544	51.95622019	95.01056348	142.6029013	Agriculture	1800	0
A	N	2	S	M	17.04792857	21	50.60603455	92.6482028	130.7190885	315.3852101	Forest	1700	1
A	M	1	S	S	45.86340333	22	66.40407862	84.32338904	120.5508115	227.8247747	Urban	1700	1
A	N	1	S	W	20.32952687	24	71.42226038	40.82473324	27.98550883	241.0086699	Wetland	1600	0
M	N	1	S	S	7.164072935	22	28.59200298	54.90257562	133.8838498	378.8882802	Urban	1800	0
A	N	2	S	M	35.73594168	25	19.06604449	52.86438898	80.90133629	140.9591778	Forest	1900	0
L	C	1	S	N	14.66767107	24	72.21236559	56.24837968	121.1160233	159.0198058	Urban	1700	1
H	N	1	S	N	26.27351537	21	76.89560488	54.85358622	134.413695	462.3208956	Urban	1600	0

D	N	4	S	N	34.89163928	24	70.34893321	43.75157774	47.70052125	28.11214743	Forest	1600	0
L	N	5	V	N	45.01320556	23	66.66612523	67.536431	116.5077887	242.8171492	Urban	1900	1
L	N	5	V	N	39.60955974	21	55.86693864	83.96400849	34.19027438	456.4004002	Urban	1600	0
L	N	5	V	N	33.81805339	23	48.56001594	76.40390348	64.06616829	305.1815039	Forest	1700	0
L	N	5	V	N	33.98165912	22	16.08907319	80.37230012	122.7022149	273.2035371	Agriculture	1600	0
L	N	5	V	N	47.2962322	23	38.59172491	44.86895071	129.1095875	124.8720896	Forest	1800	0
L	N	5	V	N	14.79120487	22	36.08819619	97.08944261	1.04281958	171.6377445	Urban	1800	0
M	N	1	S	N	0.05560471	22	68.18276486	90.30951003	117.6120954	426.0242079	Wetland	1700	1
L	N	2	M	N	13.59772797	23	60.63940796	88.30541911	62.61165047	279.6090659	Wetland	1800	0
L	N	2	V	N	10.89456898	24	44.71130146	89.3790283	33.31617157	257.3710969	Forest	1900	0
L	N	2	V	N	33.09172926	24	15.67323511	95.96263049	17.9798051	49.78985677	Wetland	1600	0
M	N	3	M	W	31.68557999	24	25.41282414	72.65523797	50.64227571	8.014078228	Urban	1700	0
M	N	1	S	N	29.67306914	25	57.82811346	52.01692107	141.4364556	481.8425943	Urban	1900	0
M	N	1	S	N	0.81217216	25	55.32916016	77.00701647	148.4804398	288.4930836	Forest	1600	1
L	N	2	V	N	36.43890272	25	69.58448398	84.57288962	137.8185933	301.606531	Urban	1900	1
M	N	1	S	N	16.17652062	25	44.66025689	84.2750436	105.4528438	39.85753659	Agriculture	1700	0
L	N	1	S	N	33.25279945	25	43.64106041	71.28694145	54.54444036	341.5571429	Urban	1700	0
M	N	1	S	N	27.82892143	25	51.46854493	44.10752316	145.7673124	466.2833544	Forest	1900	0
M	N	2	V	N	17.13654627	24	67.72766761	62.26677631	144.3670942	287.8326265	Agriculture	1600	1
A	N	4	M	M	6.728813355	22	34.34664455	95.24599651	37.76734437	63.05766119	Agriculture	1800	0
D	N	5	V	N	4.720944883	24	57.46113068	75.06693459	134.5872759	290.8607637	Wetland	1800	1
M	N	1	S	N	41.57591666	24	49.60123748	72.2999166	45.13174647	383.1736066	Urban	1900	0
D	N	1	S	N	45.9198771	23	58.69197891	56.13121086	142.7260742	303.2082683	Forest	1700	1
D	N	1	S	N	32.48940234	25	71.50409904	62.10693519	105.5330421	439.918874	Wetland	1600	1
M	F	4	V	W	5.171705743	24	65.81982151	93.72076535	141.4346501	324.3414767	Wetland	1600	1
M	C	2	M	S	20.10136084	25	56.09162843	79.98778285	75.40185348	449.6244591	Urban	1700	0
M	F	4	V	W	36.44696001	25	69.54072104	87.24322161	7.721812687	311.7526556	Urban	1800	0
M	N	2	S	W	38.9910442	25	70.71059407	67.26072538	41.79696964	65.643986	Forest	1600	0
M	N	1	S	S	5.910931115	25	59.58540837	77.81483317	136.2398829	354.8458556	Wetland	1900	1
M	D	4	V	N	0.001535942	25	68.59093299	54.90301652	35.9342836	100.8974741	Forest	1800	0
D	N	1	S	N	35.60685229	25	58.82300232	82.32764982	121.7342308	454.4132489	Wetland	1900	1
M	V	1	S	N	17.8298088	25	57.60985402	65.65608002	137.417914	380.2000336	Wetland	1800	1
M	N	1	S	S	12.72406658	23	53.30279648	66.55273089	147.8475681	338.1991845	Wetland	1700	0
A	N	1	S	M	0.644873489	25	62.69016477	78.95933947	136.3082907	350.4831598	Urban	1900	1
A	N	2	V	M	27.01497034	24	21.10235737	96.17683729	100.8203321	92.02238603	Wetland	1900	0
A	N	1	S	M	42.55753677	24	71.66095314	43.8403808	114.2429423	378.225544	Wetland	1700	0
A	N	2	V	W	47.88158231	24	71.02904694	89.48459424	35.6456316	237.0789729	Wetland	1900	0
D	N	3	V	S	28.28724753	24	12.04730981	57.54298662	109.2324523	112.8947761	Agriculture	1800	0
M	N	3	V	W	25.72306639	23	67.80717254	66.63516727	135.1674699	308.3109891	Urban	1600	1
A	F	3	V	S	4.252819274	24	19.02089072	41.31479973	94.84587459	20.16692607	Agriculture	1600	0
L	F	1	S	N	27.44364008	24	33.4583198	58.06276467	95.02945661	163.0851271	Urban	1700	0
M	V	1	M	W	18.99895395	24	62.04557794	70.15787456	80.36620261	234.2784973	Forest	1800	0

A	V	2	S	N	30.35676404	25	51.25319272	43.37056464	113.5434655	373.9791671	Agriculture	1900	1
A	A	2	V	S	19.43631739	25	67.25769169	69.46574376	125.2953743	492.3272388	Forest	1900	1
M	N	2	V	S	12.01495617	26	68.24939246	95.62663773	148.1170097	104.5486823	Wetland	1600	1
M	N	1	M	N	4.748263761	23	45.52274136	46.32359309	27.97777656	65.17656502	Forest	1900	0
M	N	1	S	S	15.74592223	25	10.44701102	85.86644368	6.116271233	101.7674369	Forest	1600	0
M	N	2	S	S	4.832241714	26	30.09266932	64.5802632	88.63394148	384.5305666	Forest	1700	0
A	N	1	S	N	8.844847541	24	53.18488429	79.31042287	101.6346543	212.7528296	Forest	1900	0
A	N	2	V	W	49.36680793	23	78.68303246	55.61419588	121.4881743	282.7079699	Urban	1800	1
A	N	2	M	W	22.20517785	25	54.22694689	49.56953575	76.81395874	10.50457849	Agriculture	1900	0
A	V	2	S	S	26.59358833	25	68.18625067	49.62775686	133.9743663	417.2156543	Urban	1600	1
M	N	2	M	M	43.68092644	25	54.38039922	44.22951674	96.77591856	222.9127749	Urban	1700	0
M	N	2	S	S	49.77844502	25	47.79897658	51.13880229	26.15496435	408.3276216	Urban	1600	0
A	N	1	S	W	29.13572247	24	64.58917766	79.8531631	103.6406607	442.6194394	Forest	1900	1
A	C	5	V	S	40.62369105	24	77.48864472	92.90143671	148.0103019	443.3532838	Urban	1600	1
M	N	2	S	N	16.34705122	24	63.27195318	88.84754114	140.5094983	269.1815511	Agriculture	1600	0
A	N	3	M	S	15.28719787	24	74.88866051	81.10805859	120.6281416	456.6148431	Urban	1700	1
H	N	3	V	M	20.13518605	26	77.40944027	46.6259118	51.15995266	193.9863772	Wetland	1800	0
L	N	1	V	S	33.62536313	26	73.93105162	57.35124678	117.0210282	414.7829157	Forest	1900	1
M	N	1	S	M	34.10958446	26	54.28353252	58.58841252	138.7040427	366.8478272	Wetland	1700	1
M	N	4	V	N	15.76197173	22	75.24196739	54.99741678	131.600903	467.6972448	Agriculture	1800	1
M	N	3	S	W	6.660496207	23	17.17568096	70.90033958	38.69124416	375.7116588	Agriculture	1600	0
M	N	4	V	M	31.61981282	23	75.60599411	72.13322889	138.9976069	470.6671964	Agriculture	1800	1
M	N	2	S	N	6.416293109	23	58.15200056	61.41315851	122.58333	338.3086168	Agriculture	1600	0
M	F	3	M	S	28.92462541	20	14.74859414	61.22734952	83.28012174	181.701966	Urban	1600	0
A	N	2	S	W	34.6711064	21	31.06744969	89.71333587	139.4475868	469.0905863	Forest	1700	1
A	N	2	S	W	35.03057482	23	59.57204621	87.3575709	36.27784364	137.7833882	Agriculture	1600	0
A	N	1	S	W	37.67360553	20	14.7145421	58.4677499	13.96541517	166.086261	Wetland	1700	0
A	N	1	S	N	43.64297108	20	50.75193221	94.82730663	134.5823637	350.404209	Agriculture	1800	0
A	N	1	S	W	24.9979496	22	34.21181399	97.16890102	135.0627086	382.6581498	Forest	1800	0
M	N	2	S	W	36.51245948	30	53.46408624	59.60455252	94.96521859	464.7017419	Wetland	1600	1
M	N	2	S	W	30.95863677	25	63.20194237	61.26549693	125.8544687	302.4012045	Wetland	1700	1
A	N	2	S	M	9.33199699	28	71.00757643	70.33804921	135.3814362	398.9990176	Wetland	1600	1
A	N	2	S	W	1.284482243	22	78.14422784	96.46724875	108.8933518	369.2730267	Urban	1700	0
A	N	2	S	S	14.22706928	20	77.82144987	92.57916599	134.566539	311.8972781	Agriculture	1700	1
A	N	2	V	M	22.15168532	21	62.47562822	46.15407776	143.0629636	293.6115706	Urban	1700	1
A	N	2	V	N	30.85529722	28	69.10603681	63.5638446	126.9813319	341.3827285	Wetland	1800	1