**Table of Contents**

[**1 Introduction** 4](#_Toc158794626)

[**1.1 Research Objectives** 5](#_Toc158794627)

[**1.1.1 Research Problem** 5](#_Toc158794628)

[**1.1.2 Hypothesis** 6](#_Toc158794629)

[**1.1.3 Research Objectives** 6](#_Toc158794630)

[**2. Literature Review** 7](#_Toc158794631)

[**2.1 Introduction** 7](#_Toc158794632)

[**2.2 Traditional Fire Risk Assessment Methods** 8](#_Toc158794633)

[**2.3 Data Science Applications in Environmental Risk Assessment** 10](#_Toc158794634)

[**2.4 Comparative Analysis of Fire Risk Assessment in Different Geographies** 11](#_Toc158794635)

[**2.5 Advancements in Predictive Analytics and Modelling for Fire Risk Assessment** 12](#_Toc158794636)

[**2.6 Economic and Social Implications of Fire Risk in Irish CCF Systems** 14](#_Toc158794637)

[**2.7 Policy and Regulatory Frameworks Impacting Fire Management in Irish CCF** 15](#_Toc158794638)

[**2.8 Gaps and Limitations in Current Research** 15](#_Toc158794639)

[**2.9 Justification for the Present Study** 17](#_Toc158794640)

[**3. Methodology:** 19](#_Toc158794641)

[**3.1 Data Collection:** 23](#_Toc158794642)

[**3.1.1 CD's API Data:** 23](#_Toc158794643)

[**3.1.2 Simulated CCF Forest Data Variables:** 24](#_Toc158794644)

[**3.2 Data Cleaning:** 26](#_Toc158794645)

[**3.3 Label Encoding:** 27](#_Toc158794646)

[**3.4 Handling Imbalanced Data:** 28](#_Toc158794647)

[**3.5 Exploratory Data Analysis:** 30](#_Toc158794648)

[**3.5.1 Scatter Plots for Feature Analysis:** 30](#_Toc158794649)

[3.5.2 Visualization of Categorical Features: 30](#_Toc158794650)

[**3.6 Feature Engineering:** 32](#_Toc158794651)

[**3.7 Data Scaling and Splitting:** 34](#_Toc158794652)

[**3.7.1 Standard Scaling with StandardScaler:** 35](#_Toc158794653)

[3.7.2 Dataset Splitting into Training and Testing Sets: 35](#_Toc158794654)

[**3.8 Tools Used:** 36](#_Toc158794655)

[**4 Implementation** 36](#_Toc158794656)

[**4.1 KNN Classification** 36](#_Toc158794657)

[**4.1.1 Explanation:** 37](#_Toc158794658)

[**4.1.2 Parameters Used** 38](#_Toc158794659)

[**4.2 X\_Gradient Boosting:** 38](#_Toc158794660)

[**4.2.1 Explanation:** 38](#_Toc158794661)

[**4.2.2 Parameters:** 39](#_Toc158794662)

[**4.3 Random Forest Classifier** 39](#_Toc158794663)

[**4.3.1 Explanation:** 40](#_Toc158794664)

[**4.3.2 Parameters:** 40](#_Toc158794665)

[**4.4 Decision Tree Classifier** 41](#_Toc158794666)

[**4.4.1 Explanation:** 41](#_Toc158794667)

[**4.4.2 Performance Note:** 42](#_Toc158794668)

[**4.5 Support Vector Machine (SVM) Regression** 43](#_Toc158794669)

[**4.5.1 Explanation:** 43](#_Toc158794670)

[**4.5.2 Parameters:** 44](#_Toc158794671)

[**5 Evaluations** 45](#_Toc158794672)

[**5.1 Model Evaluation for KNN Classification:** 45](#_Toc158794673)

[**5.1.1 Model Performance** 46](#_Toc158794674)

[**5.1.2 Visualizations:** 46](#_Toc158794675)

[**5.2 Model Evaluation for XGBoost Classifier:** 47](#_Toc158794676)

[**5.2.1 Evaluation Metrics:** 47](#_Toc158794677)

[**5.2.2 Visualizations:** 48](#_Toc158794678)

[**5.3 Model Evaluation for Random Forest Classifier:** 49](#_Toc158794679)

[**5.3.1 Evaluation Metrics:** 49](#_Toc158794680)

[**5.3.2 Visualization:** 49](#_Toc158794681)

[**5.4 Model Evaluation for Decision Tree Classifier:** 50](#_Toc158794682)

[**5.4.1 Evaluation Metrics:** 50](#_Toc158794683)

[**5.4.2 Visualization:** 51](#_Toc158794684)

[**5.5 Model Evaluation for Support Vector Machine (SVM) Regression:** 51](#_Toc158794685)

[**5.5.1 Evaluation Metrics:** 51](#_Toc158794686)

[**5.5.2 Visualization:** 52](#_Toc158794687)

[**6 Results & Conclusions:** 53](#_Toc158794688)

[**6.1 Accuracy comparison difference** 53](#_Toc158794689)

[**6.2 Observations:** 53](#_Toc158794690)

[**6.3 Model Selection for Deployment:** 55](#_Toc158794691)

[**6.4 Dashboard:** 58](#_Toc158794692)

[**6.5 Conclusion:** 59](#_Toc158794693)

**Table Of Figures**

[**Figure 1 Methodology 20**](#_Toc158794748)

[**Figure 2 Data Flow Diagram 21**](#_Toc158794749)

[**Figure 3 Bar Graph 31**](#_Toc158794750)

[**Figure 4 Bar Graph 31**](#_Toc158794751)

[**Figure 5 Line Graph 32**](#_Toc158794752)

[**Figure 6 Correlation matrix 33**](#_Toc158794753)

[**Figure 7 Updated Correlation matrix 34**](#_Toc158794754)

[**Figure 8 classification report KNN 45**](#_Toc158794755)

[**Figure 9 Confusion Matrix KNN 47**](#_Toc158794756)

[**Figure 10 classification report XG Boost 48**](#_Toc158794757)

[**Figure 11 Confusion Matrix XG Boost 48**](#_Toc158794758)

[**Figure 12 classification report RFC 49**](#_Toc158794759)

[**Figure 13 Confusion Matrix RFC 50**](#_Toc158794760)

[**Figure 14 classification report Decision Tree 50**](#_Toc158794761)

[**Figure 15 Confusion Matrix Decision tree 51**](#_Toc158794762)

[**Figure 16 classification report SVM 52**](#_Toc158794763)

[**Figure 17 Confusion matrix SVM 52**](#_Toc158794764)

[**Figure 18 Accuracy Comparison 54**](#_Toc158794765)

[**Figure 19 Overall Comparison 54**](#_Toc158794766)

[**Figure 20 Deployment Diagram 55**](#_Toc158794767)

[**Figure 21 Web App 58**](#_Toc158794768)

**Abstract**

**Add In once results achieved**

**Keywords:** Continuous Cover Forestry, Data Science, Fire Risk Prediction, Machine Learning, Climate Change, Sustainable Forestry Management.

# **1 Introduction**

This research initiative is placed at the intersection of ecological conservation and technological innovation, seeking to transform fire risk assessment in Irish Continuous Cover Forestry (CCF) through the application of advanced data science techniques. The study will analyse a wealth of available data sources, such as detailed satellite imagery, sequential meteorological data, thorough soil studies, and extensive documentation of forest management techniques, to develop an all-encompassing predictive model. This model is anticipated to significantly enhance the capabilities for forecasting fire incidents and formulating effective prevention and response strategies in the context of CCF. Continuous Cover Forestry represents a shift in forest management, advocating for the maintenance of an uneven-aged and species-diverse forest canopy that mimics natural woodland conditions. This approach promotes a sustained forest ecosystem, which is recognized for its multitude of environmental benefits such as biodiversity conservation, soil stabilization, water regulation, and carbon sequestration (O'Hara, 2014). Moreover, CCF is acknowledged for fostering greater forest resilience to pests, diseases, and climate change-induced stressors (Pommerening and Murphy, 2004). Despite its advantages, the application of CCF in Ireland introduces complexities in fire risk assessment due to the variable forest structures and the diverse species assemblies that characterize these forests. Conventional fire risk models, typically designed for uniform, even-aged plantations, fall short when applied to the heterogeneous and dynamic environments of CCF (Fernandes and Botelho, 2003). The diverse canopy layers, varied ground vegetation, and the presence of continuous cover create a unique fire environment where traditional prediction models can neither accurately capture the fuel load distributions nor the potential fire behaviour. This study intends to fill the gap by leveraging data science to dissect and understand the distinctive fuel matrices and fire dynamics of CCF. Spatial analytics and machine learning algorithms will be employed to analyse the spatial distribution of various forest stands, the accumulation of biomass, and the moisture content in different forest layers—factors that are critical to predicting fire ignition and spread. Historical fire data and simulations will be used to train and validate the model, ensuring it reflects the complexities of CCF. In addition, the research will explore the impact of climatic variables such as temperature fluctuations, precipitation patterns, and wind dynamics on fire risk within CCF. With climate change exacerbating the frequency and intensity of forest fires globally (IPCC, 2021), this investigation is paramount for a comprehensive understanding of the evolving fire risk landscape. The aim is to create a dynamic model that not only considers the current state of the forest but also incorporates projections of how changing climatic conditions could influence future fire risks. By developing a tool that is both scientifically rigorous and practical for on-the-ground forest management, this research will contribute significantly to the safeguarding of Ireland's forest ecosystems. Furthermore, the insights gained from this study will have broader implications, potentially guiding fire risk assessment and forest management strategies in other regions where CCF is practiced or being considered. This research aspires to inform policy, improve forest health and resilience, and ensure that the ecological, economic, and social benefits of forests are protected against the rising threat of wildfires.

## **1.1 Research Objectives**

### **1.1.1 Research Problem**

The central research problem of this project is the absence of a sophisticated, data-driven methodology for assessing and predicting fire risk within Irish Continuous Cover Forestry. As CCF becomes more prevalent as a sustainable forest management practice, the development of precise risk assessment tools becomes imperative to prevent forest fires, particularly as such events may be intensified by climate change and the complex structures of diverse forests.

### **1.1.2 Hypothesis**

Our guiding hypothesis posits that a data science-driven predictive model, employing advanced machine learning algorithms, will set a new standard in fire risk assessment for CCF—an area where established predictive systems are currently lacking. This innovative model will leverage a comprehensive dataset, amalgamating environmental, meteorological, and forest management variables, to unearth complex patterns and interactions that evade traditional methods, thereby significantly improving the accuracy of fire risk predictions within CCF landscapes.

## **1.1.3 Research Objectives**

To construct a predictive model for fire risk in CCF, we set forth the following objectives:

1. Comprehensive Methodology Review: To perform an exhaustive review and critique of current fire risk assessment methods. This will identify critical criteria and variables specifically pertinent to fire risk in CCF settings, such as fuel loads, tree species diversity, and canopy structures, informing the development of a superior model that addresses the shortcomings of existing approaches.
2. Development of Predictive Model: To create a sophisticated, data-driven predictive model using machine learning techniques such as Random Forests or Gradient Boosting Machines. This model will incorporate a wide array of variables—spanning climatic conditions, topography, vegetation characteristics, and historical fire occurrences—to generate nuanced and precise fire risk forecasts tailored to the unique ecosystem dynamics of CCF.
3. Model Validation Framework: To validate the predictive model's accuracy, a robust framework will be employed, utilizing historical fire incidence data alongside rigorous statistical analyses. We will apply cross-validation techniques, delve into error metrics like Root Mean Square Error (RMSE), and conduct sensitivity analyses to confirm the model's performance under various conditions. This meticulous process will refine the model's predictive power, cementing its role as a dependable tool for CCF fire risk management.
4. Decision-Support Tool Development: To transform the predictive model into a user-friendly decision-support tool designed for forest managers and policymakers. This tool will integrate the model's advanced analytics into an interactive interface, presenting actionable insights that facilitate informed decision-making for fire risk mitigation in CCF landscapes. We will illustrate the tool's application through scenario-based simulations to demonstrate its practical utility in real-world settings.

These objectives are intricately linked to the proposed title "Leveraging Data Science for Fire Risk Assessment in Irish Continuous Cover Forestry." The objectives set a comprehensive plan for creating, validating, and implementing a data science model to fulfil the pressing need for advanced fire risk assessment tools in CCF practices, as identified by the research problem and hypothesis.

# **2. Literature Review**

## **2.1 Introduction**

The significance of accurately assessing fire risk within Continuous Cover Forestry (CCF) systems is critical for the sustainability of forest management and environmental conservation. This literature review systematically examines the breadth of scholarly research to discover current knowledge and pinpoint gaps specifically in fire risk assessment, emphasizing the integration of data science techniques. Amidst the growing prevalence of forest fires and the exacerbating effects of climate change, the pertinence of this review is clear; it lays the groundwork for an empirical study aimed at bolstering fire risk prediction methodologies in Irish CCF environments (O’Sullivan et al., 2017; Forest Service, 2021). The review is organized thematically, focusing on the tenets of CCF, the evolution from traditional to modern fire risk assessment approaches, the incorporation of data science in environmental risk analysis, and a comparative evaluation of fire risk assessment practices across various geographic regions. This thematic structure enables the construction of an integrated perspective on the status quo of fire risk assessment approaches and the feasibility of sophisticated predictive frameworks in forest management. Literature was meticulously selected based on criteria that prioritize peer-reviewed articles, authoritative governmental reports, and significant scientific publications from the past two decades, with consideration for seminal works that constitute the foundation of the field. Selection criteria based on relevance to CCF practices, fire risk assessment proficiency, data science applications within environmental disciplines, and case studies from Ireland as well as other pertinent regions. Academic databases such as Scopus, ScienceDirect, Google Scholar, and the Forestry Commission's research archives were systematically searched, guaranteeing a comprehensive collection of insights ranging from regional Irish contexts to broader international research endeavours (McCarthy et al., 2006; FAO, 2018). Continuous Cover Forestry represents a paradigm shift in forest management, promoting the conservation of semi-natural habitats through selective harvesting while preserving structural diversity and forest ecology (Pommerening and Murphy, 2004). By avoiding clear-cutting, CCF aims to create a heterogeneous forest environment with a natural age structure and species variety, contributing to increased ecosystem resilience (O’Hara, 2016). Core to the philosophy of CCF are principles such as uninterrupted forest cover, the fostering of native species, and reduced anthropogenic disturbances, aligning closely with the dynamics of natural forests (Mason et al., 1999). Fire risk in CCF environments warrants particular attention due to the complex forest architecture, which can lead to unpredictable wildfire behaviour. Variables such as mixed age stands, species diversity, and the accumulation of deadwood create a fuel landscape that deviates significantly from uniform forestry stands, affecting both fire propagation and intensity (Lindenmayer et al., 2012).

Therefore, the development of sophisticated risk models is essential for forecasting fires within CCF, enabling proactive management and minimizing ecological disruption (Rodríguez et al., 2014). Fire, while a natural ecological agent, poses increased risks in proximity to human development and under changing climatic conditions, with global and local trends indicating rising fire frequency and severity (Bowman et al., 2009; Moritz et al., 2014). Although the history of forest fires in Ireland has been associated with traditional forestry, the emergence of CCF necessitates a revised understanding of fire dynamics within these alternative forestry practices (Forest Service Ireland, 2020). In summary, the ecological benefits of CCF are accompanied by distinct fire management challenges. The historical context of wildfires, both internationally and within Ireland’s conventional forestry, informs the development of fire risk strategies. The transition to CCF thus calls for innovative, context-specific approaches to fire management that harmonize the protection of ecosystems with fire risk mitigation (FAO, 2018)

## **2.2 Traditional Fire Risk Assessment Methods**

Fire risk assessment in forestry has traditionally relied on a combination of empirical data, statistical models, and expert knowledge. Conventional methods often use historical fire data to identify patterns and factors most frequently associated with the initiation and spread of wildfires. These methods have been integrated into various fire danger rating systems globally, like the Canadian Forest Fire Danger Rating System (CFFDRS) and the National Fire Danger Rating System (NFDRS) used in the United States (Stocks et al., 1989; Deeming et al., 1977) and the McArthur Forest Fire Danger Index (FFDI) utilized in Australia Fire risk assessment tools usually include factors such, as weather conditions (like temperature, wind speed, humidity) topography and vegetation types to determine a fire danger index. It is evident that there is a requirement for the creation of tools and approaches that can overcome the restrictions of fire risk assessments and offer dependable information, for managing CCF in Ireland (Thompson and Calkin 2011). In more traditional forestry practices that are characterized by uniform stand structures and species compositions, such methods have been effective in predicting areas of high fire risk and in guiding fire prevention efforts (Andrews et al., 2003). However, when evaluating their effectiveness in Continuous Cover Forestry, the limitations of traditional fire risk assessment methods become apparent. These methos assume that the fuel is uniformly distributed and behaves in a predictable way, which is not the case in CCF (Zumbrunnen et al., 2011). Furthermore, the microclimate variations within CCF stands, a result of the canopy complexity, can significantly influence fore behaviour in ways that traditional models may not accurately capture (Agee, 1993). Another limitation is the temporal resolution of traditional assessment methods. While these methods are suitable for short-term fire risk predictions, they are less effective for long-term risk assessments, which are essential for planning in the context of CCF where the forest structure is managed over extended periods (Fernandes and Botelho, 2003). Challenges also arise from the scale at which these models operate. Traditional methods often work at larger scales and may not provide the fine-scale resolution necessary for managing CCF plots, where individual tree selection and small-scale challenges, particularly in regions where fire occurrences are low or where fire suppression efforts have been highly effective, resulting in a lack of data to inform risk models. This data limitation is further compounded in CCF systems, where historical fire data may be non—representative of current conditions due to changes in forest management practices over time (McCarthy et al., 2001). While traditional fire risk assessment methods have provided a foundation for understanding and managing wildfire risk in conventional forestry practices, their limitations are pronounced in the context off CCF. The complexity and variability inherent in CCF require more nuanced and dynamic modelling approaches that can accommodate the unique features of these forests.

## **2.3 Data Science Applications in Environmental Risk Assessment**

Data science has notably transformed environmental risk assessment by enriching the analysis, prediction, and management processes with advanced technological tools. Leveraging big data analytics, machine learning (ML), and spatial analysis has unlocked deeper insights into complex environmental risks. The multifaceted approach of data science in this field encompasses diverse methodologies. For instance, machine learning algorithms have significantly enhanced the prediction of forest fire likelihood through the examination of extensive datasets comprising weather variables, vegetation types, and historical fire instances (Rodrigues and de la Riva, 2014). Techniques like artificial neural networks (ANNs) and support vector machines (SVMs) offer a rapid and precise assessment of fire risks, outpacing conventional statistical methods (Viegas et al., 2019). A case in point is the application of data science in assessing bushfire risks in Australia, where the integration of machine learning with remote sensing has facilitated the mapping of fuel loads and fire behaviour prediction across diverse terrains (Pettinari and Chuvieco, 2016). Furthermore, satellite imagery and advanced data processing tools have become instrumental for real-time fire monitoring and for strategic deployment of firefighting resources. In the United States, projects like Monitoring Trends in Burn Severity (MTBS) employ satellite remote sensing technology to evaluate the severity and extent of burns, which enriches the understanding of fire effects on ecosystems and supports the rehabilitation of areas impacted by fires (Eidenshink et al., 2007). Innovations in geographic information systems (GIS) have enabled the layering and spatial analysis of various data sets, culminating in sophisticated fire susceptibility models that integrate topographical, meteorological, and human factors to present a holistic view of landscape-level fire risks (Chuvieco et al., 2010). Additionally, the Internet of Things (IoT) has revolutionized risk assessments, with sensor networks in forests delivering real-time environmental data to facilitate dynamic risk modeling responsive to changes in factors like temperature and humidity (Garcia-Chevesich et al., 2017). Despite these advancements, the field faces considerable challenges, particularly in the realms of data volume and computation. Accurate predictions require significant computational power and algorithms designed for efficient data processing (Sakr et al., 2011). Moreover, the validity of these models hinges on the quality and granularity of the data, which can be variable based on the source. Data science has made a profound impact on environmental risk assessment. The sophistication and efficacy of these tools have not only improved the accuracy and response times of analyses but are also set to become more central in environmental risk management, especially regarding forest fires and similar natural threats.

## **2.4 Comparative Analysis of Fire Risk Assessment in Different Geographies**

Fire risk assessment strategies vary globally, reflecting the diverse climatic conditions, forest management practices, and technological capabilities of different regions. This section explores these strategies with a particular focus on countries like Ireland, Canada, Australia, and other European nations, and how local conditions influence fire risk assessment. In Ireland, fire risk assessment for forestry has historically been less of a concern compared to countries with a more prevalent fire regime. However, climate change and shifts in land use are altering fire dynamics, necessitating a revaluation of risk assessment strategies (Luo et al., 2014). The Irish strategy now includes using meteorological data to forecast fire weather, incorporating these data into fire danger rating systems adapted from models like the Canadian Forest Fire Weather Index (FWI) System (Davies and Gray, 2015). Canada, with its vast boreal forests, is well-acquainted with wildfires, and the country has developed one of the most advanced fire danger systems in the world. The Canadian FWI system is a comprehensive tool that uses weather data to estimate fire danger and has been adapted by many countries due to its reliability (Wotton et a., 2009). Its ability to account for regional differences unweather and vegetation makes it a versatile tool for fire risk assessment Similar to Canada, Australia faces significant wildfire threats across its varied landscapes, from tropical savannas to temperate eucalypt forests known for their flammability. The country has developed its fire danger rating systems, which include the McArthur Fire Danger Index and the Australian Fire Danger Rating System, using local climate data, vegetation types, and fire behaviour research. These systems are essential for providing fire danger warning to both the public and fire management authorities in a country that experiences frequent and sever bushfires (Tian et al., 2005). In Europe, countries experience a diverse set of fire environments, from the Mediterranean's fire-prone ecosystems to the less fire-affected forests of central and northern Europe. The European Forest Fire Information System (EFFIS) aids fire management by integrating satellite data and ground-based observations to provide near-real-time monitoring of fire events (San-Miguel-Ayanz et al., 2012), playing a vital role in harmonizing fire risk assessment across the continent's diverse landscapes.

The impact of climatic differences on fire risk assessment is critical. The dry, hot summers of Mediterranean Europe and Australia's bushfire-prone regions create conditions conducive to wildfires, in stark contrast to the cooler, wetter climate of northern Europe and Ireland, resulting in lower fire incidence. These climatic factors, along with geographical variations such as topography and road networks, significantly influence fire behaviour and risk assessment (Modugno et al., 2016). Management practices also play a significant role in shaping fire risk assessments. For instance, the traditional 'fight fire with fire' techniques employed by some Canadian First Nations and Australian Aboriginal cultures, which use controlled burns to manage the landscape, contrast with the fire suppression policies historically prevalent in European forestry (Lewis et al., 2018). Fire risk assessment strategies must be tailored to local conditions, incorporating climatic, geographical, and management factors. By learning from the varied approaches of countries like Ireland, Canada, Australia, and those in Europe, stakeholders can develop more effective, context-specific risk assessments. The adaptive capacity of fire risk assessment systems, such as the Canadian FWI, Australia’s national frameworks, and Europe's EFFIS, serves as models for countries worldwide as they confront the evolving challenges posed by climate change on fire regimes.

## **2.5 Advancements in Predictive Analytics and Modelling for Fire Risk Assessment**

Advancements in predictive analytics have become a cornerstone for fire risk assessment, particularly with the fusion of machine learning and environmental science. The transformative impact of deep learning (DL), such as the implementation of Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), is evident in their capacity to process and analyse the sequential and temporal dynamics of fire-related datasets, leading to nuanced modelling of fire risk progression over time (Goodfellow et al., 2016). Ensemble models, employing both bagging and boosting techniques, have significantly improved prediction accuracy by mitigating the variance and bias commonly associated with more simplistic models, thereby delivering a robust predictive performance (Breiman, 2001). Additionally, time series forecasting methods, including Autoregressive Integrated Moving Average (ARIMA) models, leverage historical fire occurrence and climatic data to forecast future fire risks with enhanced precision (Box et al., 2015). The integration of big data in fire risk analytics has offered unprecedented scope and scale in monitoring and assessment. Satellite imagery from MODIS and VIIRS has been instrumental in providing daily updates on surface temperatures and vegetation health, key factors in determining fire risk (Justice et al., 2002). Concurrently, the deployment of Internet of Things (IoT) technologies has facilitated the real-time acquisition of crucial environmental parameters, such as soil moisture levels and canopy temperatures, directly impacting the precision of fire potential assessments (Ray, 2016). Real-time data processing, facilitated by edge computing, marks a pivotal enhancement in fire management strategies. By processing sensor and satellite data at or near the source, edge computing dramatically reduces latency, thereby supporting swift and informed decision-making during critical fire episodes (Shi et al., 2016). Furthermore, streaming analytics platforms can digest and interpret data streams in real time, offering immediate insights into fire behaviour and enhancing the efficiency of early warning systems (Zaharia et al., 2013). Noteworthy case studies exemplifying the use of advanced analytics in fire risk assessments include the AI-driven prediction models employed during California's Camp Fire in 2018, which utilized live data streams for predicting the fire's trajectory, thus facilitating timely evacuations (Kaggle, 2018). Similarly, in Brazil, fire risk models incorporating climate data and utilizing Support Vector Machines (SVMs) have shown substantial success in predicting fire incidences with high accuracy (Maeda et al., 2011). Despite these advancements, the integration of sophisticated predictive models in fire risk assessment is not without challenges. The validity and detail of data can influence the accuracy of predictions, while the computational intensity required for processing extensive datasets and executing intricate simulations necessitates robust computational infrastructure (Dean et al., 2008). Moreover, there exists a critical need for interdisciplinary expertise to effectively translate complex model outputs into actionable fire management strategies, necessitating collaboration between data scientists and fire management experts (Kolden et al., 2015). Looking to the future, predictive analytics in fire risk management is expected to evolve towards the utilization of AI with Unmanned Aerial Vehicles (UAVs) for meticulous data collection and nuanced risk assessments (Yuan et al., 2015). The prospect of developing autonomous fire detection systems employing satellite and drone technologies promises a leap forward in rapid response capabilities. Furthermore, integrating AI with comprehensive climate models can provide valuable foresight into the long-term effects of climate change on fire risk, enabling more effective preparation and mitigation strategies (Amatulli et al., 2013). This holistic approach to predictive analytics signifies a profound shift in fire risk assessment, propelling it towards a future where data-driven insights drive proactive, rather than reactive, fire management strategies. These advancements collectively represent a step change in our capability to predict, manage, and reduce the risks associated with wildfires, showing the potential of technology to safeguard natural resources and communities alike.

## **2.6 Economic and Social Implications of Fire Risk in Irish CCF Systems**

In Ireland, the adoption of Continuous Cover Forestry systems carries distinct economic and social implications, particularly when considering the risk of wildfires. Economically, fire incidents within CCF landscapes can incur substantial costs, including those associated with challenging suppression efforts due to the complex forest structures promoted by CCF principles. Moreover, such fires threaten significant revenue losses from the destruction of commercially valuable timber, adversely affecting the livelihoods of those dependent on forestry in rural areas **(Ní Dhubháin et al., 2023).** The potential for damage to CCF areas also poses threats beyond immediate timber losses. It compromises the broader ecological benefits, such as carbon sequestration, which is crucial for Ireland's climate action targets **(Ní Dhubháin et al., 2023).** Additionally, wildfires can have detrimental effects on tourism, which is vital to the rural economy, as areas of natural beauty and recreational forests are rendered inaccessible or less appealing post-fire (Prestemon et al., 2016). Socially, the ramifications extend to public health concerns, notably when peatlands, which are prevalent in Irish landscapes, catch fire. The smoke and particulate matter from burning peat have been associated with respiratory problems and other health risks, necessitating attention in fire management policies (Rein et al., 2008). Though large-scale evacuations are less common in Ireland due to the smaller size of forested areas, the social fabric of rural communities can still be disrupted by fireincidents, affecting residents' sense of security and community cohesion **(Ní Dhubháin et al., 2023)**. In conclusion, Ireland’s CCF systems must navigate the fine balance between sustainable forest management and the heightened risk of wildfires. The economic and social stakes highlight the need for tailored fire management strategies that align with the intricate dynamics of CCF, thereby safeguarding not only the forests but also the communities that surround and depend upon them.

Is this needed?

## **2.7 Policy and Regulatory Frameworks Impacting Fire Management in Irish CCF**

In Ireland, the policy and regulatory landscape is a fundamental component that dictates fire management approaches within Continuous Cover Forestry (CCF) systems. The Forest Service, operating under the Department of Agriculture, Food, and the Marine, oversees efforts to prevent and control forest fires, with regulatory power granted by key legislation such as the Forest Act and the Wildlife Acts. These statutes outline the legal parameters for conducting prescribed burns and establish punitive measures for the initiation of illegal fires (Forest Service, 2017). The incorporation of predictive analytics into fire risk assessment is in part shaped by these regulatory frameworks. Initiatives like the National Forest Fire Danger Rating system illustrate the government's efforts to embed predictive tools within public policy. Nevertheless, the practical application of such tools often encounters obstacles, such as the necessity for current fire management strategies to evolve to fully exploit sophisticated predictive analytics. This need for evolution must also consider compliance with broader European Union environmental directives, which aim to harmonize environmental protection measures across member states (Davies et al., 2016). To harness the full potential of these technologies, regulatory frameworks need to be dynamic and forward-looking. The Forestry Programme 2014-2020, for instance, has laid down provisions to bolster forest protection that could be further developed to embrace advanced analytics and other contemporary fire management technologies (Forest Service, 2020). Effective integration also hinges on cross-departmental and interagency collaboration, which is crucial to crafting comprehensive policies that support the application of the best available technology within CCF practices (Fischer et al., 2016). In embracing advanced predictive analytics through deliberate policy measures and regulatory frameworks, Ireland can significantly enhance the capacity of its CCF systems to withstand the increasing wildfire threat, while also contributing to global best practices in sustainable forest management.

## **2.8 Gaps and Limitations in Current Research**

Despite the significant advancements in fire risk assessment within Continuous Cover Forestry (CCF), the literature reveals persistent gaps and limitations that need to be addressed to enhance predictive capabilities and management interventions. One of the notable gaps is the limited research on the specific application of fire risk assessment tools in CCF systems. Most fire risk models have been developed for and tested within conventional forestry settings, where clear cutting and even-aged stand structures predominate. CCF's complex forest structures, with their continuous canopy and multi-aged stands, present unique fire behaviour that is not adequately captured by these models (Pukkala, 2018). This highlights the need for developing or adapting models that can account for the intricacies of CCF. Another gap is the integration of climate change projections into fire risk assessment. While climate models predict increased fire weather severity, integrating these projections into risk models remains challenging due to the uncertainties associated with climate models and the long-term nature of forest planning (Johnston et al., 2016). The slow incorporation of this data undermines the potential for current assessments to anticipate future fire regimes accurately. The variability in the scale of data collection also presents a limitation. Much of the forest fire data are collected at a macro scale, which can dilute the specific conditions encountered in CCF areas (Bowman et al., 2011). The lack of fine-scale, high-resolution data impedes the ability to conduct nuanced analyses that are essential for CCF areas, which are heterogeneous by nature. Furthermore, there is a scarcity of longitudinal studies that evaluate the long-term effectiveness of fire risk reduction strategies in CCF contexts. Such studies are critical for understanding the temporal dynamics of fire risk and the sustained impacts of management interventions (Thompson and Calkin, 2011). Without this information, it is challenging to develop adaptive management strategies that evolve with changing forest conditions and risk profiles. The literature also points to limited understanding of socio-economic dimensions of fire risk in CCF. While ecological and climatic factors have received considerable attention, the human dimension, including public perception, economic impacts, and policy considerations, is less well-studies within the context of CCF (Fischer, 20011). As such, there is a need for holistic approaches that incorporate these factors into fire risk assessment and management. In terms of methodological limitations, there is a reliance on historical fire data to predict future risks. Given the changing climate and land-use patterns., past fire regimes may not be indicative of future conditions (Mortiz et al., 2012). Therefore, there is a need for forward-looking approaches that cab adapt to changing environmental and human landscapes. Lastly, while data science techniques offer promising avenues for enhancing fire risk assessment, the integration of these techniques into operational settings is still in its infancy (Rodrigues and de la Riva, 2014). Challenges such as data availability, computational resources, and the need for interdisciplinary expertise hinder the operationalization of these advanced methods. In summary, the current body of literature on fire risk assessment in CCF is burgeoning, yet there are significant gaps and limitations that must be addressed. The development of models specific to CCF, incorporation of climate change projections, collection of fine-scale data, longitudinal studies, consideration of socio-economic factors, forward-looking predictive methods, and operational integration of data science techniques are all areas that require further research and attention.

## **2.9 Justification for the Present Study**

The escalating complexities of forest fire management, particularly within the domain of Continuous Cover Forestry (CCF) in Ireland, necessitate a revaluation and enhancement of existing fire risk assessment methods. The current literature underscores a clear need for advanced approaches that can encapsulate the nuanced ecological dynamics of CCF systems (O'Sullivan et al., 2019). Irish CCF, with its distinctive landscape and climatic conditions, presents a unique set of challenges that standard fire risk models, primarily developed for more homogeneous forest structures, struggle to address effectively (Murphy et al., 2016). The potential of data science to revolutionise this field is considerable. Data science, with its capacity to process large and complex datasets, offers novel insights into fire behaviour enabling more accurate risk predictions (Joyce and Rodman, 2018). By integrating heterogenous data sources, such as satellite imagery, weather patterns, and topography, alongside algorithms capable of learning from this data, the development of a predictive model specific to the Irish context is within reach (Smith et al., 2016). Such a model would not only augment the precision of risk assessments but also serve as a pivotal tool for forest management and policymaking. The intersection of data science and fire risk assessment in Irish CCF is aligned with the research questions posed by this study, probing the effectiveness of current methods and the feasibility of implementing a data-driven approach. It addresses the gaps highlighted in previous sections, such as the need for a tailored models for CCF and the incorporation of fine scale, high resolution data into fire risk assessment. (Fernandes and Botelho, 2019). Moreover, it responds to the call for innovative strategy that can integrate evolving climate change projections and socio economies factors into forest management practice (McCarthy et al. 2018). This study is poised to make a significant contribution to the field by providing empirical evidence on the applicability and benefits of data science techniques in CCF fire risk assessment. It aims to bridge the current research gaps by operationalizing a sophisticated, data centric, approach, potentially setting a benchmark for future research and practice in the realm of sustainable forestry management (Davies et al., 2020).  
Fire risk assessment is a crucial component of sustainable forest management, especially in the context of climate change and increasing fire frequency and severity (FAO, 2019). Fire is a natural disturbance that shapes forest ecosystems and influences their structure, composition, and function (Bond et al., 2005). However, fire can also have negative impacts on forest resources, such as timber production, carbon sequestration, biodiversity conservation and human safety (FAO, 2019). Therefore, it is important to understand the factors that affect fire risk and behaviour, and to develop appropriate strategies and tools to prevent, control and mitigate fire effects.

Forests play a pivotal and indispensable role on a global scale, serving as vital contributors to biodiversity conservation, carbon sequestration, and the provision of ecosystem services. In Ireland, most forests are even aged monoculture forests dominated by Sitka spruce (Picea sitchensis) plantations, which constitute 52% of the planted species. Lodgepole pine (Pinus contorta) accounts for 12% while other conifers and broadleaves contribute around 9% and 27% respectively (Teagasc, 2021). An Increase in public awareness over loss of biodiversity and limited ecosystem benefits associated with Sitka spruce plantations has gained attention (DAFM, 2014). The forest cover in Ireland is low at 11% of the land area. Among these forests, 52% are under state ownership managed by Coillte while the remaining 48% are privately owned (Department of Agriculture, Food, and the Marine, 2020). Despite efforts to encourage afforestation the average annual rate from 2016 to 2020 fell short of the government’s target of 8,000 hectares, reaching 4,200 hectares (Teagasc, 2021). Consequently, in response to these circumstances, there has been a shift in forest policy towards promoting multifunctional forests that can meet multiple objectives such as timber production, biodiversity conservation and recreation. One such approach is continuous cover forestry (CCF), which is an alternative silvicultural system that aims to maintain a permanent forest cover and enhance biodiversity, resilience, and ecosystem services (O'Hara, 2014; DAFM, 2014). CCF is based on the principles of natural forest dynamics, selective harvesting, and continuous regeneration. CCF has been promoted as a more sustainable and adaptive form of forestry in the face of climate change and social demands (Pommerening et al., 2015). CCF aims to enhance the ecological and social benefits of forests, as well as their resilience to climate change (Pommerening et al., 2015). However, one of the challenges of CCF is to manage the fire risk associated with complex forest structures and diverse species compositions. There is a scarcity of data and tools to assess and mitigate fire risk in CCF systems, and more research is needed to support fire-adapted management practices (Ní Dhubháin et al., 2019).

This paper presents a proof of concept for a data-driven approach to fire risk assessment in Irish CCF forests, using machine learning techniques to model the relationship between forest structure, fuel characteristics and fire behaviour. Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions without explicit programming (Mitchell, 1997). Machine learning has been widely applied to various fields and problems, including fire science and management (Rodrigues et al., 2020). The paper reviews the existing literature on fire risk in CCF, describes the data collection and analysis methods, and discusses the preliminary results and implications for CCF management in Ireland.

The paper is structured as follows. Section 2 provides a brief overview of the concepts and definitions of fire risk, fire behaviour and CCF. Section 3 reviews the current state of knowledge on fire risk in Forestry systems, with a focus on European studies. Section 4 describes the data sources, variables and methods used for the machine learning analysis. Section 5 presents the results of the analysis, including the performance evaluation and interpretation of the models. Section 6 discusses the main findings and limitations of the study, as well as the potential applications and benefits of the proposed approach for CCF management. Section 7 concludes with some recommendations for future research and development.

**Impact of Climate Change and Land-Use Change on Wildfires: A UN Report**

**The increasing frequency and intensity of wildfires globally are at the forefront of environmental concerns, as outlined in a recent report by the UN Environment Programme (UNEP) and GRID-Arendal. The report, titled Spreading like Wildfire: The Rising Threat of Extraordinary Landscape Fires, emphasizes the escalating risk of wildfires due to climate change and land-use alterations. Projections indicate a potential surge of up to 14% in extreme fires by 2030, 30% by 2050, and a staggering 50% by the end of the century.**

**Rewrite this section- add in more algorthim info**

**Reference list summary lit review table**

# **3. Methodology:**

Every Supervised learning Project has an overall methodology like:

A diagram of a machine learning algorithm

Description automatically generated

Figure 1 Methodology

The Process includes:

* Data Extraction
* Data Cleaning
* Label Encoding
* Handling Imbalanced Data
* Exploratory Data Analysis
* Feature Engineering
* Data Splitting and Scaling
* ML Algorithms
* Results Evaluations  
  Each rectangle is representing a process and the relation between each process is given as process flow(arrow is representing the flow). We will discuss each process in detail in the next sections.

**Data Flow Diagram**

The process flow diagram involves the several processes given in fig 2.

A diagram of a data flow

Description automatically generated

Figure 2 Data Flow Diagram

Explanation

**User:**

Explanation: The process begins with the user, who initiates the entire workflow. The user may be involved in specifying requirements, providing input data, or interacting with the system in many ways.

**Datasets:**

Explanation: The Datasets step involves gathering and preparing the data required for the forest fire risk assessment. This could include historical weather data, simulated forest data, or any other relevant information needed to train and evaluate the machine learning model.

**Preprocessing:**

Explanation: In the Preprocessing step, the collected datasets undergo data cleaning, transformation, and feature engineering. This ensures that the data is in a suitable format for training the machine learning model. Preprocessing may involve handling missing values, scaling features, or encoding categorical variables.

**Model Training:**

Explanation: Model Training is where the machine learning model is developed using the pre-processed datasets. The model learns patterns and relationships from the data, allowing it to make predictions about forest fire risk based on input features.

**Evaluation:**

Explanation: After training, the model's performance is evaluated using a separate dataset not seen during training. Evaluation metrics such as accuracy, precision, recall, or F1 score are used to assess how well the model generalizes to new, unseen data. The Evaluation step is connected to the Evaluation Results for a comprehensive understanding of the model's performance.

**Model Selection:**

Explanation: In Model Selection, different machine learning models or variations of the same model are compared based on their performance during evaluation. The goal is to choose the most effective model for predicting forest fire risk.

**Deployment:**

Explanation: Once a suitable model is selected, it is deployed for use in real-world scenarios. Deployment involves integrating the model into a system or application, making it accessible for generating predictions based on new input data.

**Real World Predictions:**

Explanation: The deployed model is now capable of making real-world predictions. It takes input data relevant to forest conditions, weather, or other parameters and produces predictions about the likelihood of a forest fire occurring.

**End:**

Explanation: The process concludes with the End step, signifying the completion of the workflow. Users can now leverage the forest fire risk assessment system to obtain predictions and insights based on the deployed machine learning model. The interconnected flow ensures a seamless transition between each stage, enhancing the overall effectiveness of the forest fire risk assessment system.

## **3.1 Data Collection:**

In this Project we must use two dataset’s details of the datasets are given below:

### **3.1.1 CD's API Data:**

The dataset obtained from the Copernicus Emergency Management Service provides a comprehensive historical reconstruction of meteorological conditions conducive to the initiation, propagation, and sustainability of fires. This dataset is an integral component of the European Forest Fire Information System (EFFIS) and encompasses fire danger metrics based on models developed in Canada, the United States, and Australia. The fire danger indices within this dataset are calculated using weather forecasts derived from historical simulations provided by the ECMWF ERA5 reanalysis. ERA5, a globally complete and consistent dataset, combines model data with quality-controlled observations, making it a reliable proxy for observed atmospheric conditions. The dataset includes variables such as the Build-Up Index, Burning Index, Danger Rating, Drought Code, Energy Release Component, Fine Fuel Moisture Code, and various others. These variables play crucial roles in assessing fire danger levels and formulating fire danger forecasts for pre-suppression planning.

**Explanation of Variables Related to Weather from CD's API:**

The CD's API dataset comprises various key variables related to weather, each playing a distinctive role in assessing fire danger levels and predicting fire behaviour. Here is a summary of the significant variables:

* Build-Up Index (Dimensionless): A weighted combination of the Duff Moisture Code and Drought Code, indicating the total fuel available for combustion. Often used for pre-suppression planning.
* Burning Index (Dimensionless): Measures the difficulty of controlling a fire, derived from the Spread component and Energy Release component.
* Danger Rating (Dimensionless): Equivalent to the Fire Weather Index (FWI), reduced to 6 classes of danger, providing a harmonized spatial distribution of fire danger levels.
* Drought Code (Dimensionless): Indicates moisture content in deep compact organic layers, representing a fuel layer at 10-20 cm deep.
* Drought Factor: Represents fuel availability, given as a number between 0 and 10, influenced by recent temperatures and rainfall events.
* Duff Moisture Code (Dimensionless): Indicates moisture content in loosely compacted organic layers of moderate depth, representative of the duff layer that is 5-10 cm deep.
* Energy Release Component (J/m2): Reflects the potential heat release per unit area in the flaming zone, providing guidance on fire intensity.
* Fine Fuel Moisture Code (Dimensionless): Indicates moisture content in litter and other cured fine fuels, with a scale ranging from 0-99.
* Fire Daily Severity Index (Dimensionless): Numeric rating of the difficulty of controlling fires, exponentially increasing as the Fire Weather Index rises.
* Fire Danger Index (Dimensionless): Metric related to the chances of a fire starting, its rate of spread, intensity, and difficulty of suppression.
* Fire Weather Index (Dimensionless): A combination of Initial Spread Index and Build-Up Index, indicating the potential frontal fire intensity.
* Ignition Component (%): Measures the probability a firebrand will require suppression action, expressed as a probability on a scale of 0 to 100.
* Initial Spread Index (Dimensionless): Combines the Fine Fuel Moisture Code and wind speed to indicate the expected rate of fire spread.
* Keetch-Byram Drought Index (Dimensionless): Represents cumulative moisture deficiency in deep duff and upper soil layers, providing insight into flammability.
* Spread Component (Dimensionless): Measures the theoretical ideal rate of fire spread, expressed as a dimensionless variable.

### **3.1.2 Simulated CCF Forest Data Variables:**

The simulated Continuous Cover Forestry (CCF) forest data provides essential variables that contribute to assessing the overall fire risk in Ireland. Each variable represents distinct aspects of the forest environment and conditions, collectively offering valuable insights for understanding and predicting fire behaviour. Here's a summary of the significant variables:

* Overall Fire Risk (Numeric): Represents the comprehensive assessment of fire risk in the simulated CCF forest, derived from a combination of various contributing factors.
* Fine Fuel Moisture (Numeric): Indicates the moisture content in litter and other cured fine fuels, influencing the flammability of surface-level vegetation.
* Initial Spread Index (Numeric): Reflects the expected rate of fire spread based on the combination of fine fuel moisture and wind speed in the simulated CCF forest.
* Uneven-Aged Canopy (Binary): Indicates the presence or absence of an uneven-aged canopy structure, influencing the spatial distribution of vegetation and potential fire spread.
* Species Diversity (Numeric): Represents the variety of tree species within the simulated CCF forest, influencing overall ecosystem resilience and fire dynamics.
* Continuous Canopy Cover (Binary): Indicates whether the forest has continuous canopy cover, impacting fuel continuity and potential fire spread.
* Drought Conditions (Numeric): Reflects the simulated moisture deficiency in the forest, providing insights into the potential flammability of organic material in the ground.
* Wind Speed (Numeric): Represents the simulated wind speed within the forest environment, a critical factor influencing fire behaviour and spread.
* Temperature (Numeric): Indicates the simulated temperature conditions within the CCF forest, influencing overall fuel moisture and fire risk.
* Fire Warnings (Binary): Indicates whether there are simulated fire warnings in the CCF forest, providing information on potential fire risk events.
* Fire Occurrence (Binary): Indicates the occurrence or absence of simulated fires in the CCF forest, serving as a crucial outcome variable for assessing the accuracy of fire risk models.

These variables collectively contribute to a comprehensive understanding of fire risk dynamics in the simulated CCF forest. The combination of meteorological data from CD's API and these simulated forest-specific variables forms a robust foundation for developing accurate and effective machine learning models for forest fire risk assessment in Ireland.

## **3.2 Data Cleaning:**

Data cleaning is a critical process in the data preparation phase that focuses on enhancing the quality, accuracy, and reliability of a dataset. It involves various techniques to identify and rectify issues such as missing values, outliers, inconsistencies, and irrelevant information.  
Following techniques are applied for Data Cleaning:

**Handling Missing Values:**

Technique: Imputation or removal of missing values.

Description: Fill missing values using methods such as mean, median, or regression for numerical variables. Alternatively, remove records or variables with missing values.

**Outlier Detection and Treatment:**

Technique: Statistical methods (Z-score, IQR) for identifying outliers, and removal or transformation.

Description: Identify outliers using statistical measures and address them by either removing them or transforming them to minimize their impact.

**Consistency Check:**

Technique: Comparison of values against predefined rules or standards.

Description: Ensure consistency across variables by validating values against predefined criteria or rules.

**Normalization/Scaling:**

Technique: Min-Max scaling or Z-score normalization for numerical variables.

Description: Normalize numerical features to a similar scale, preventing features with larger magnitudes from dominating the analysis.

**Validation of Binary Variables:**

Technique: Validation to ensure binary variables only contain valid values (0 or 1).

Description: Confirm that binary variables adhere to the expected format, containing only valid values.

**Data Integration:**

Technique: Integration of datasets based on common identifiers (e.g., time periods, geographic locations).

Description: Combine datasets with common identifiers, creating a unified dataset for analysis.

**Data Type Check:**

Technique: Inspection and correction of data types for each variable.

Description: Confirm that variables are appropriately represented as numerical or categorical types.

**Data Duplicates:**

Technique: Identification and removal of duplicate records.

Description: Detect and remove duplicate records to eliminate redundancy in the dataset.

**Data Format Standardization:**

Technique: Standardizing date formats and other data representations.

Description: Ensure consistent data formats, such as date formats, to maintain uniformity across the dataset.

**Handling Irrelevant Variables:**

Technique: Removal of irrelevant or redundant variables.

Description: Eliminate variables that do not contribute significantly to the analysis, streamlining the dataset.

## **3.3 Label Encoding:**

In the label encoding section, the 'OverallFireRisk' column has undergone the following mapping:

Mapping for 'OverallFireRisk' column:

Extreme: 0

High: 1

Low: 2

Moderate: 3

Very Low: 4

This label encoding technique transforms categorical descriptors of fire danger levels into unique numerical labels. The assigned numeric values, ranging from 0 to 4, allow machine learning algorithms to comprehend and analyse the fire risk data effectively. This process enhances the readiness of the data for subsequent stages in the forest fire risk assessment, as machine learning models often require numerical representations for accurate training and prediction. Label encoding provides a streamlined approach to convert qualitative information into a format compatible with machine learning algorithms, contributing to the overall success of the forest fire risk assessment endeavour.

## **3.4 Handling Imbalanced Data:**

Dealing with Imbalanced Data in CCF Forest Fire Risk Assessment Model:

**Background:**

Imbalanced data occurs when certain classes in the target variable are underrepresented. In the context of CCF forest fire risk assessment, the target variable is 'OverallFireRisk,' which encompasses various levels of fire risk.

**Imbalanced Class Distribution:**

'OverallFireRisk' exhibits an imbalanced distribution, where specific fire risk levels may have fewer instances compared to others. This imbalance can lead to challenges in model training, with the risk of the model being biased towards the majority class.

**Significance of Imbalanced Data Handling:**

Addressing the imbalanced class distribution is crucial to ensure that the machine learning model can effectively learn patterns and make accurate predictions for all fire risk levels. This is particularly important in scenarios where accurate identification of minority classes is essential for decision-making.

**Target Variable Overview:**

'OverallFireRisk' consists of categories representing diverse fire risk levels, such as 'Extreme,' 'High,' 'Moderate,' 'Low,' and 'Very Low.' Each category signifies a specific degree of fire risk associated with CCF forests.

**RandomOverSampler Technique:**

To tackle the imbalanced class distribution, the code incorporates the RandomOverSampler technique from the imbalanced-learn library. RandomOverSampler is an oversampling method designed to address the scarcity of instances in the minority class.

**Mechanism of RandomOverSampler:**

RandomOverSampler works by randomly duplicating instances of the minority class until a more balanced distribution is achieved. This process involves creating synthetic samples for the less frequent class, mitigating the impact of class imbalance during model training.

**Prevention of Biased Model Training:**

The primary goal of using RandomOverSampler is to prevent machine learning models from disproportionately favoring the majority class during training. This helps in creating a more equitable learning environment, ensuring that the model considers all fire risk levels.

**Enhanced Generalization:**

The application of RandomOverSampler enhances the model's ability to generalize across all classes of fire risk. This is crucial for achieving a well-rounded and unbiased prediction capability, especially when dealing with imbalanced datasets.

**Contribution to Improved Model Performance:**

The resulting balanced dataset, generated through RandomOverSampler, significantly contributes to improved model performance. It enables the model to provide more accurate predictions for all fire risk levels, thereby enhancing its overall reliability and effectiveness.

**Practical Relevance:**

In real-world scenarios, where imbalanced classes could have profound consequences, the use of RandomOverSampler ensures that the developed CCF forest fire risk assessment model is robust and capable of handling varying frequencies within the 'OverallFireRisk' variable.

## **3.5 Exploratory Data Analysis:**

Exploratory Data Analysis is a crucial phase in the data science pipeline, offering insights into the dataset's characteristics, relationships between variables, and patterns that can guide subsequent modeling decisions. In this context, the EDA process focuses on understanding the features and their impact on the target variable 'OverallFireRisk.'

### **3.5.1 Scatter Plots for Feature Analysis:**

Scatter plots are employed to visually inspect the relationship between various features and the target variable 'OverallFireRisk.' For each feature, a scatter plot is generated, with the x-axis representing the feature's values and the y-axis denoting the corresponding 'OverallFireRisk' levels. By examining these plots, trends, patterns, and potential outliers can be identified. This aids in understanding how individual features contribute to the variability in fire risk levels.

## 3.5.2 Visualization of Categorical Features:

* Bar Graphs:

Categorical features, such as 'drtcode' and 'FireWarnings,' are visually explored using bar graphs. Bar graphs provide a clear representation of the distribution of 'OverallFireRisk' across distinct categories within each categorical feature. This enables the identification of any significant variations in fire risk levels associated with specific categories. See fig 3 & 4.

A bar graph with blue and black lines

Description automatically generated

Figure 3 Bar Graph

A bar graph with blue squares

Description automatically generated

Figure 4 Bar Graph

* Line Graphs:

Line graphs are utilized to illustrate the impact of categorical features on 'OverallFireRisk' over a continuous range. This visualization method helps uncover trends or patterns in fire risk variation concerning the values of categorical features. See fig 5.

A barcode with text on it

Description automatically generated

Figure 5 Line Graph

## **3.6 Feature Engineering:**

In the feature engineering phase of this project, strategic decisions were made to enhance the quality and relevance of the dataset, 'balanced\_data,' for subsequent machine learning model training. The initial step involved a comprehensive analysis of the dataset's correlation matrix, utilizing the seaborn library to create informative heatmaps. This analysis aimed to identify inter-feature relationships and dependencies, guiding the selection of features for further refinement.

Upon assessing the correlation matrix, a judicious approach to feature selection was employed. Several features were deemed redundant or exhibited high correlation with others, potentially introducing multicollinearity issues. To address this, specific features were identified for removal, including 'surface,' 'FireWarnings,' 'ffmcode,' 'fdsrte,' 'dufmcode,' 'fwinx,' 'Unnamed: 0,' 'time,' 'fdimrk,' 'drtcode,' 'FireOccurrence,' and 'fbupinx.' The removal of these features was executed to streamline the dataset, eliminating redundancies, and reducing the dimensionality of the feature space.

Following the feature selection and removal process, an updated correlation matrix was generated and visualized using a heatmap. This provided a visual representation of the refined feature relationships within the modified dataset. The correlation matrix showcased the impact of feature engineering on mitigating multicollinearity and optimizing the dataset for the subsequent machine learning tasks.

The culmination of these feature engineering efforts ensures that the 'balanced\_data' dataset is not only more computationally efficient but also poised to contribute meaningfully to the training of machine learning models for fire risk assessment. The judicious selection and refinement of features lay a solid foundation for improved model interpretability, generalization, and predictive accuracy.

A screenshot of a data analysis

Description automatically generated

Figure 6 Correlation matrix

A screenshot of a computer screen

Description automatically generated

Figure 7 Updated Correlation matrix

## **3.7 Data Scaling and Splitting:**

Data scaling and splitting are essential preprocessing steps in preparing the dataset for machine learning model training and evaluation. This process ensures that the features are on a similar scale, preventing certain features from dominating others and facilitating better convergence during model training. Additionally, splitting the dataset into training and testing sets allows for an unbiased evaluation of the model's performance.

### **3.7.1 Standard Scaling with StandardScaler:**

Standard scaling, or z-score normalization, is employed to bring numeric features to a standard scale with a mean of 0 and a standard deviation of 1. This is crucial when features have different units or scales, ensuring that each feature contributes equally to model training. The StandardScaler from the scikit-learn library is applied to achieve this normalization.

**Detailed Steps:**

* Import the StandardScaler from scikit-learn's preprocessing module.
* Identify the numeric features in the dataset that require scaling.
* Initialize an instance of the StandardScaler.
* Fit the scaler on the training data to compute the mean and standard deviation.
* Transform both the training and testing sets using the computed mean and standard deviation.

The scaling process ensures that the features are centred around zero, providing a standardized input for the machine learning models. This step is particularly important for algorithms sensitive to the scale of features, such as K-Nearest Neighbors or Support Vector Machines.

### 3.7.2 Dataset Splitting into Training and Testing Sets:

The dataset is divided into two subsets: the training set and the testing set. The training set is used to train the machine learning models, while the testing set is reserved for evaluating the models' performance on unseen data. The train\_test\_split function from scikit-learn is commonly used for this purpose.

**Detailed Steps:**

* Import the train\_test\_split function from scikit-learn's model\_selection module.
* Identify the features (X) and the target variable (Z) in the dataset.
* Specify the test\_size parameter to determine the proportion of data allocated to the testing set (commonly set to 0.2 for an 80-20 split).
* Set the random\_state parameter for reproducibility of results.
* Execute the train\_test\_split function to obtain the training and testing sets for both features and the target variable.

The resulting sets, namely X\_train, X\_test, Z\_train, and Z\_test, are ready for use in training and evaluating machine learning models. This separation ensures that the model is assessed on data it has never seen during the training phase, providing a reliable estimate of its generalization performance.

By meticulously performing data scaling and splitting, the code establishes a robust foundation for subsequent model training, enhancing the models' ability to generalize to new, unseen data and improving the overall reliability of the machine learning pipeline.

## **3.8 Tools Used:**

* **CDS API**: Used for Scrapping the data.
* **Language:** Python
* **Pandas:** Used for data manipulation and reading CSV files.
* **Scikit-learn:** Employed for label encoding, data scaling, dataset splitting, and implementing machine learning algorithms.
* **Imbalanced-learn:** Utilized for addressing class imbalance using RandomOverSampler.
* **Matplotlib and seaborn:** Used for data visualization, including scatter plots, bar graphs, line graphs, and heatmaps.
* **GridSearchCV:** Applied for hyperparameter tuning of machine learning models.
* **Front-end Development:** HTML CSS
* **Back-end development:** Python Django

# **4 Implementation**

## **4.1 KNN Classification**

In KNN classification, the problem is formulated as a classification task aiming to predict the overall fire risk. K-Nearest Neighbors (KNN) is employed as a classification algorithm. The model calculates the distances between data points and predicts the target variable based on many k-nearest neighbors. This method is suitable for capturing local patterns and relationships within the dataset.

### **4.1.1 Explanation:**

The application of the K-Nearest Neighbors (KNN) algorithm to the fire risk assessment dataset proved to be a strategic choice, emphasizing proximity-based classification. The initial step involved an exploration of hyperparameter tuning, leveraging GridSearchCV to systematically evaluate various combinations of hyperparameters. The parameters considered included the number of neighbors ('n\_neighbors'), the weighting strategy ('weights'), and the distance metric ('p'). This meticulous tuning aimed to identify the optimal configuration that maximizes the model's predictive accuracy.

Hyperparameter tuning is crucial as it helps fine-tune the algorithm's behavior, optimizing its performance on the specific dataset. The selection of the best hyperparameters, as determined by GridSearchCV, ensures that the KNN model is tailored to the unique characteristics of the fire risk assessment data.

To assess the model's generalization capabilities and mitigate overfitting concerns, a robust cross-validation strategy was employed. The cross\_val\_score function executed a 5-fold cross-validation, which involves splitting the dataset into five subsets, training the model on four subsets, and evaluating it on the fifth. This process was repeated five times, with each subset serving as the validation set exactly once. Cross-validation provides a more comprehensive evaluation of the model's performance across different subsets of the data, offering insights into its stability and consistency.

The hyperparameter-tuned KNN classifier demonstrated exceptional performance on the test set, achieving an accuracy of 100%. The precision, recall, and F1-score metrics, as presented in the classification report, all reached perfect scores, underscoring the model's ability to precisely predict fire risk categories. This outstanding performance validates the effectiveness of hyperparameter tuning and cross-validation, highlighting their significance in ensuring a well-optimized and robust KNN model for fire risk assessment.

### **4.1.2 Parameters Used**

Optimal number of neighbors are 3, 5, 7,9. The number of neighbors used in KNN is determined through hyperparameter tuning to find the optimal value that maximizes the model's accuracy. This process involves trying different values for the number of neighbors and evaluating the model's performance with each setting. The justification for selecting a specific number of neighbors is to strike a balance between model complexity and generalization.

* Fewer Neighbors (e.g., 3 or 5): Using a small number of neighbors can make the model sensitive to noise or outliers in the data. While it may capture local patterns well, it might also be influenced by individual data points, leading to overfitting.
* More Neighbors (e.g., 7 or 9): On the other hand, using a larger number of neighbors can lead to a smoother decision boundary, potentially improving the model's ability to generalize to unseen data. However, it might overlook local patterns and details in the data.

The optimal number of neighbors is often chosen based on cross-validated performance metrics, such as accuracy, to ensure that the model performs well on new, unseen data. The goal is to find a balance that minimizes both bias and variance, resulting in a KNN model that generalizes effectively to different instances in the dataset. The choice of the number of neighbors is data-dependent and may vary based on the characteristics of the dataset and the underlying patterns in the data.

## **4.2 X\_Gradient Boosting:**

Gradient Boosting is applied using the XGBoost algorithm, treating the task as a classification problem. XGBoost is an ensemble technique that combines weak learners to build a robust model. It iteratively builds decision trees and optimizes for both bias and variance, resulting in improved accuracy. The algorithm is trained to classify the fire risk into distinct categories based on the features provided.

### **4.2.1 Explanation:**

The application of XGBoost, an advanced implementation of gradient boosting, to the fire risk assessment dataset signifies a sophisticated modeling approach, leveraging the strengths of an ensemble technique. The process commenced with an intricate exploration of hyperparameter tuning, employing GridSearchCV to systematically evaluate a range of hyperparameter combinations. Key parameters under consideration included the number of trees ('n\_estimators'), maximum depth of the trees ('max\_depth'), learning rate ('learning\_rate'), and subsample ratio of training instances ('subsample'). This exhaustive search aimed to identify the most effective hyperparameter configuration for optimizing the model's accuracy.

### **4.2.2 Parameters:**

Hyperparameter tuning is imperative in the context of XGBoost due to its sensitivity to parameter values, and optimal settings can significantly impact the model's performance. The selected hyperparameters, derived from GridSearchCV, ensure the XGBoost model is tailored to the unique characteristics of the fire risk assessment dataset.

To comprehensively assess the model's generalization performance and robustness, a cross-validation strategy was implemented. Cross-validation, particularly the 5-fold approach in this case, provides a more nuanced evaluation of the model's consistency across various subsets of the data. This approach enhances the model's reliability and ensures it can generalize well to unseen data.

The hyperparameter-tuned XGBoost classifier demonstrated robust performance on the test set, achieving an accuracy of approximately 93.7%. The classification report further details precision, recall, and F1-score metrics, highlighting the model's ability to effectively classify fire risk categories. This successful application of hyperparameter tuning and cross-validation underscores their importance in refining and evaluating the XGBoost model, contributing to its effectiveness in predicting fire risk levels accurately.

## **4.3 Random Forest Classifier**

The Random Forest Classifier is utilized as an ensemble learning method to address the classification task. This technique creates multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is effective in handling complex relationships and reducing overfitting.

### **4.3.1 Explanation:**

The application of the Random Forest classifier to the fire risk assessment dataset represents a robust ensemble learning technique that leverages multiple decision trees to make accurate predictions. The initial step involved a meticulous exploration of hyperparameter tuning, employing GridSearchCV to systematically evaluate various hyperparameter combinations. Parameters such as the number of trees ('n\_estimators'), maximum depth of the trees ('max\_depth'), minimum samples required to split an internal node ('min\_samples\_split'), and minimum samples required to be a leaf node ('min\_samples\_leaf') were fine-tuned to identify the configuration that maximized the model's accuracy.

### **4.3.2 Parameters:**

The selected hyperparameters for the Random Forest classifier were determined through a process of hyperparameter tuning using GridSearchCV. Let's delve into the technical reasons behind why these specific hyperparameters were chosen over others:

**Number of Estimators (n\_estimators):**

Contribution: The parameter dictates the quantity of decision trees in the ensemble, impacting the model's overall robustness.

Justification: The grid search systematically explored various values for n\_estimators, ensuring an optimal balance between improved model stability and computational efficiency.

**Maximum Depth of the Trees (max\_depth):**

Contribution: Regulating the depth of individual trees mitigates the risk of overfitting, particularly as deeper trees can capture intricate patterns in the training data.

Justification: The grid search sought the ideal max\_depth by assessing its impact on the trade-off between capturing complex relationships and preventing overfitting.

**Minimum Samples Split (min\_samples\_split) and Minimum Samples Leaf (min\_samples\_leaf):**

Contribution: These parameters control node splitting and leaf formation, respectively, influencing the granularity of decision tree structures.

Justification: The grid search dynamically adjusted min\_samples\_split and min\_samples\_leaf to strike a balance, enabling the model to generalize effectively without overly specific splits tailored to the training set.

**Feature Subset Size (max\_features):**

Contribution: Randomly selecting a subset of features at each split introduces diversity among trees, reducing correlation and enhancing overall model performance.

Justification: The grid search meticulously determined the optimal max\_features, ensuring an effective compromise between diversity and predictive accuracy.

## **4.4 Decision Tree Classifier**

A Decision Tree Classifier is implemented to model the relationship between the features and the fire risk. Decision trees split the dataset based on the most significant attributes, creating a tree-like structure. The model makes predictions by traversing the tree from the root to a leaf node. It is a simple yet powerful algorithm for classification tasks.

### **4.4.1 Explanation:**

The application of the Decision Tree classifier to the fire risk assessment dataset represents a straightforward yet powerful approach to predictive modeling. The process commenced with a systematic exploration of hyperparameter tuning, utilizing GridSearchCV to evaluate a range of hyperparameter combinations. Key parameters considered for tuning included the maximum depth of the tree ('max\_depth'), the minimum number of samples required to split an internal node ('min\_samples\_split'), and the minimum number of samples required to be a leaf node ('min\_samples\_leaf'). This optimization process aimed to identify the configuration that maximizes the model's accuracy.

Hyperparameter tuning is critical in Decision Tree models to find the right balance between model complexity and overfitting. The selected hyperparameters, derived from the GridSearchCV process, ensure that the Decision Tree classifier is appropriately configured for the specific nuances of the fire risk assessment data.

To comprehensively evaluate the model's generalization capabilities and robustness, a cross-validation strategy was implemented. The 5-fold cross-validation approach allowed for a thorough examination of the model's consistency across different subsets of the data, providing insights into its stability and performance.

The hyperparameter-tuned Decision Tree classifier demonstrated moderate performance on the test set, achieving an accuracy of approximately 57.8%. The classification report further details precision, recall, and F1-score metrics, shedding light on the model's effectiveness in classifying fire risk categories. This application of hyperparameter tuning and cross-validation highlights their importance in refining and evaluating the Decision Tree model, contributing to its reliability and accuracy in predicting fire risk levels. However, it's worth noting that Decision Trees might struggle to capture complex relationships present in the data compared to more sophisticated ensemble methods.

### **4.4.2 Performance Note:**

The Decision Tree classifier, despite its simplicity and interpretability, may not perform as well as other classifiers like Random Forest or XGBoost in certain scenarios. Here are some reasons why the Decision Tree may not be performing well in this case:

**Overfitting:**

* Decision trees are prone to overfitting, especially when they are deep and capture noise in the training data.
* Overfitting occurs when the tree is too complex, capturing both the underlying patterns and the noise in the training data, leading to poor generalization to new, unseen data.

**Lack of Ensemble Effect:**

* Unlike ensemble methods such as Random Forest or XGBoost, a single decision tree may lack the robustness provided by aggregating multiple trees.
* Ensemble methods are designed to mitigate overfitting and improve overall performance by combining predictions from multiple weak learners.

**Limited Expressiveness:**

* Decision trees may struggle to capture complex relationships in the data when the relationships are not inherently hierarchical or when feature interactions are intricate.
* Other models like Random Forest or XGBoost can handle more sophisticated relationships.

**Sensitivity to Data Variations:**

* Decision trees are sensitive to variations in the training data, and slight changes in the dataset can result in different tree structures.
* This sensitivity can lead to instability in the model's performance.

**Insufficient Hyperparameter Tuning:**

The hyperparameters of the Decision Tree Classifier, such as the maximum depth and minimum samples split, might not have been fine-tuned adequately during the hyperparameter tuning process.

## **4.5 Support Vector Machine (SVM) Regression**

SVM Regression is applied to predict the area of land burnt. SVM looks for a hyperplane that best separates the data into different classes while considering outliers. The epsilon loss function is used for regression in high-dimensional space. Various kernels such as RBF, linear, polynomial, and sigmoid are explored to find the optimal hyperplane.

### **4.5.1 Explanation:**

The application of Support Vector Machines (SVM) to the fire risk assessment dataset demonstrates the versatility of this algorithm in handling both linear and non-linear classification tasks. The process commenced with an in-depth exploration of hyperparameter tuning, employing GridSearchCV to systematically evaluate various hyperparameter combinations. Key parameters considered for tuning included the regularization parameter ('C'), the kernel type ('linear', 'rbf', 'poly'), and the gamma parameter ('scale', 'auto'). This comprehensive search aimed to identify the optimal hyperparameter configuration that maximizes the model's accuracy.

Hyperparameter tuning is crucial in SVM models to find the right balance between model flexibility and regularization, ensuring optimal performance on the specific dataset. The selected hyperparameters, derived from the GridSearchCV process, tailor the SVM classifier to the unique characteristics of the fire risk assessment data.

To thoroughly assess the model's generalization capabilities and robustness, a cross-validation strategy was implemented. The 5-fold cross-validation approach facilitated a comprehensive evaluation of the model's consistency across various subsets of the data, enhancing its reliability and ability to generalize effectively. The hyperparameter-tuned SVM classifier demonstrated robust performance on the test set, achieving an accuracy of approximately 88.8%. The classification report further details precision, recall, and F1-score metrics, highlighting the model's effectiveness in accurately classifying fire risk categories. This successful application of hyperparameter tuning and cross-validation underscores their importance in refining and evaluating the SVM model, contributing to its reliability and accuracy in predicting fire risk levels. SVM's ability to handle non-linear relationships and high-dimensional data makes it a valuable tool in the context of fire risk assessments.

### **4.5.2 Parameters:**

The chosen hyperparameters for the Support Vector Machine (SVM) model are:

* **C**: 10
* **gamma**: 'scale'
* **kernel**: 'poly'

These hyperparameters contribute to the superior performance of the SVM model:

**C (Regularization Parameter):**

* C controls the trade-off between achieving a low training error and a low testing error.
* A smaller C emphasizes a smoother decision boundary, allowing for some misclassifications, while a larger C aims for a more accurate classification of each training point.
* The optimal value for C is often found through cross-validation. In the given grid, [0.1, 1, 10] represents a range of regularization strengths.

**Kernel:**

* The choice of the kernel function significantly impacts the SVM's ability to handle complex relationships in the data.
* 'Linear' kernel is suitable for linearly separable data, while 'rbf' (Radial Basis Function) and 'poly' (Polynomial) kernels can capture non-linear relationships.
* The grid ['linear', 'rbf', 'poly'] allows exploration of different kernel options.

**Gamma:**

* Gamma (γ) defines the influence of a single training example, affecting the shape of the decision boundary.
* 'Scale' uses 1 / (n\_features \* X.var()) as the gamma value, and 'auto' uses 1 / n\_features.
* Higher values of gamma lead to a more complex decision boundary, potentially causing overfitting.

# **5 Evaluations**

## **5.1 Model Evaluation for KNN Classification:**

**Evaluation Metrics:**

* Mean Absolute Error (MAE): Measures the average absolute errors between predicted and actual values.
* Mean Squared Error (MSE): Measures the average squared errors between predicted and actual values.
* R-squared: Indicates the proportion of the variance in the dependent variable that is predictable.

A screenshot of a computer

Description automatically generated

Figure 8 classification report KNN

**Cross-Validation:**

Cross-validation is performed to ensure the robustness of the KNN model.

Multiple folds are used to train and validate the model, preventing overfitting.

## **5.1.1 Model Performance**

Regarding whether the model is overfitting, based on the provided results:

* Both the training and testing accuracies are 1.0, indicating that the model achieves perfect accuracy on both datasets. The fact that both training and testing accuracies are 1.0 raises suspicion of overfitting. In a real-world scenario, achieving perfect accuracy on the testing set is uncommon and could be a sign that the model is too complex and fitting noise in the training data.
* The classification report also shows perfect precision, recall, and F1-score for all classes on the testing set. Perfect scores across all metrics in the classification report for the testing set further support the possibility of overfitting. In practical situations, a model should not perform flawlessly on all aspects, as it may indicate over-reliance on the training data's specifics.
* The training accuracy is also 1.0, suggesting that the model perfectly fits the training data. The training accuracy being 1.0 suggests that the model has memorized the training data. While high training accuracy is desirable, a perfect score might indicate overfitting, especially if the model struggles to generalize to new, unseen data.

Given these observations, there is a strong indication that the model might be overfitting. Overfitting occurs when the model captures noise or random fluctuations in the training data, leading to excessively complex decision boundaries that do not generalize well to unseen data. In this case, the model achieves perfect accuracy on both the training and testing sets, which raises suspicion of overfitting, especially in real-world scenarios where noise and variability are expected.

## **5.1.2 Visualizations:**

Visualizations include scatter plots comparing actual vs. predicted values.

Confusion metrics are visualized for different folds.

A graph with blue squares

Description automatically generated

Figure 9 Confusion Matrix KNN

## **5.2 Model Evaluation for XGBoost Classifier:**

### **5.2.1 Evaluation Metrics:**

Accuracy: The proportion of correctly classified instances.

Precision, Recall, F1-score: Metrics for assessing the performance of multi-class classification.

Hyperparameter Tuning:

GridSearchCV is used to find the best hyperparameters for the XGBoost Classifier.

A screenshot of a computer

Description automatically generated

Figure 10 classification report XG Boost

### **5.2.2 Visualizations:**

Confusion matrix heatmap: Provides a visual representation of model performance on different classes.

Bar graph comparing accuracy and other metrics.

A screenshot of a graph

Description automatically generated

Figure 11 Confusion Matrix XG Boost

**Confusion matrix of X\_gradient\_boosting:**

* True Negatives (TN): 36 (in the second row, second column)
* False Positives (FP): 1 (in the first row, second column), 1 (in the third row, second column), 2 (in the fifth row, second column)
* False Negatives (FN): 1 (in the first row, fourth column), 3 (in the second row, fourth column), 1 (in the third row, fourth column), 2 (in the fifth row, fourth column)
* True Positives (TP): 37 (in the first row, first column), 36 (in the second row, first column), 36 (in the third row, third column), 44 (in the fourth row, third column), 40 (in the fifth row, fifth column)

## **5.3 Model Evaluation for Random Forest Classifier:**

### **5.3.1 Evaluation Metrics:**

Accuracy, Precision, Recall, F1-score: Standard metrics for classification evaluation.

A screenshot of a computer

Description automatically generated

Figure 12 classification report RFC

Hyperparameter Tuning:

GridSearchCV is employed to find optimal hyperparameters for the Random Forest model.

### **5.3.2 Visualization:**

Confusion matrix heatmap: Shows classification results for different classes.

Bar graph comparing accuracy and other metrics.

A graph with numbers and squares

Description automatically generated

Figure 13 Confusion Matrix RFC

## **5.4 Model Evaluation for Decision Tree Classifier:**

### **5.4.1 Evaluation Metrics:**

Accuracy, Precision, Recall, F1-score: Standard classification metrics.

A screenshot of a computer

Description automatically generated

Figure 14 classification report Decision Tree

Hyperparameter Tuning:

GridSearchCV is used to find the best hyperparameters for the Decision Tree model.

### **5.4.2 Visualization:**

Confusion matrix heatmap: Visualizes model performance on different classes.

Bar graph comparing accuracy and other metrics.

A blue squares with white text

Description automatically generated

Figure 15 Confusion Matrix Decision tree

## **5.5 Model Evaluation for Support Vector Machine (SVM) Regression:**

## **5.5.1 Evaluation Metrics:**

Accuracy: Measures the proportion of correctly predicted instances.

R-squared: Indicates the proportion of variance in the target variable explained by the model.

A screenshot of a computer

Description automatically generated

Figure 16 classification report SVM

Hyperparameter Tuning:

GridSearchCV is employed to find the optimal hyperparameters for the SVM Regression model.

### **5.5.2 Visualization:**

Scatter plots comparing actual vs. predicted values.

Evaluation metrics visualized for different folds in cross-validation.

A screenshot of a graph

Description automatically generated

Figure 17 Confusion matrix SVM

# **6 Results & Conclusions:**

A bar graph comparing the performance metrics (accuracy, precision, recall, F1-score) of all models.

## **6.1 Accuracy comparison difference**

To compare the accuracies of different models, we can observe the differences in accuracy percentages across various machine learning models used in the project:

1. KNN:

Accuracy: 100%

1. XGBoost Classifier:

Accuracy: 93.69%

1. Random Forest Classifier:

Accuracy: 87.86%

1. Decision Tree Classifier:

Accuracy: 56.80%

1. SVM (Support Vector Machine) Classifier:

Accuracy: 88.83%

These accuracy percentages represent the overall performance of each model on the testing data. The differences in accuracy can be analyzed to understand how well each model is making correct predictions. Higher accuracy values indicate better predictive performance.

## **6.2 Observations:**

* XGBoost Classifier achieves the highest accuracy among the models, suggesting superior predictive capabilities.
* Random Forest Classifier and SVM also demonstrate high accuracy, indicating robust performance.
* KNN has a good accuracy.
* Decision Tree Classifier, while decent, falls behind the ensemble models (XGBoost and Random Forest) and SVM in terms of accuracy.

These differences in accuracy can be attributed to the inherent characteristics and strengths of each algorithm. It's important to consider factors such as model complexity, ability to handle non-linear relationships, and sensitivity to hyperparameter tuning when interpreting these accuracy values.

A table with numbers and text

Description automatically generated

Figure 18 Accuracy Comparison

Graphs for confusion matrices of classification models.

A graph of different models

Description automatically generated

Figure 19 Overall Comparison

## **6.3 Model Selection for Deployment:**

The model deployment diagram used for providing the overall web interaction is given in fig 20.

A diagram of a computer

Description automatically generated

Figure 20 Deployment Diagram

**Web Service:**

Explanation: The Web Service serves as the interface between the end user (browser or API client) and the machine learning model. It handles incoming requests, processes them, and communicates with the machine learning model to obtain predictions. This component is responsible for managing the flow of information between the user and the model.

**Machine Learning Model:**

Explanation: The Machine Learning Model is the core of the forest fire risk assessment system. It has been trained on relevant datasets to predict the risk of forest fires based on input features. These features could include weather conditions, terrain information, or other relevant parameters. The model processes incoming data from the web service and produces predictions for the likelihood of a forest fire occurring.

**Database:**

Explanation: The Database stores and manages the data required for the forest fire risk assessment. This can include historical weather data, simulated data, or any other relevant information needed for training and evaluation of the machine learning model. The model might also store and retrieve its configuration or hyperparameters from the database.

**End User (Browser or API Client):**

Explanation: The End User represents the individual or system interacting with the forest fire risk assessment application. This could be a person using a web browser to access a user interface or an external system making requests through an API. The end user provides input data (such as location, weather parameters) to the system and receives the corresponding risk assessment predictions.

**Browser or API Client:**

Explanation: The Browser or API Client is the tool used by the end user to interact with the web service. It can be a web browser for human users or an API client for automated systems. This component sends input data to the web service, receives predictions, and may display the results to the end user.

**Model Storage:**

Explanation: Model Storage is where the trained machine learning model is persisted. It could be a file storage system or a dedicated model repository. This ensures that the latest version of the model is accessible to the web service for making predictions. Storing the model separately allows for easy updates and version control.

**Deployment of KNN Model for CCF Forest Fire Risk Assessment Using Django:**

**1. Model Development:**

The machine learning model chosen for CCF forest fire risk assessment is the K-Nearest Neighbors (KNN) model. This model was trained using historical data with features relevant to CCF forests, weather conditions, and other environmental factors.

**2. Integration with Django:**

The integration with Django, a Python web framework, facilitated the development of a web application for deploying the KNN model.

**3. Frontend Development with Django Forms:**

The front-end interface was created using Django Forms, providing an intuitive and user-friendly way for end-users to input data related to CCF forests. The form captures relevant features required for the fire risk prediction.

**4. User Input Handling:**

Django manages user inputs, ensuring data integrity and preparing it for consumption by the KNN model. Data validation is implemented to guarantee that the input adheres to the expected format and range.

**5. Model Prediction Integration:**

Upon form submission, the Django application passes the entered data to the deployed KNN model for prediction. This step involves seamlessly integrating the model into the application's backend.

**6. Result Display to the User:**

The prediction generated by the KNN model is presented to the user through the Django web interface. The output can be in the form of a risk level (e.g., low, moderate, high) depending on the model output.

**7. Scalability and Accessibility:**

The Django application is designed to handle multiple users simultaneously, ensuring scalability. This enhances accessibility, allowing users to access and benefit from the model predictions concurrently.

**8. User Interaction and Iterative Input:**

Users can interact with the Django web application, providing different sets of data to obtain real-time predictions for various CCF forest scenarios. The iterative input capability enables users to explore diverse situations.

**9. User-Friendly Interface:**

The front end, built with Django templates, ensures a user-friendly experience. The interface guides users through the input process, making it straightforward for individuals with varying levels of technical expertise.

.

## **6.4 Dashboard:**

A screenshot of a computer

Description automatically generated

Figure 21 Web App

## **6.5 Conclusion:**

In conclusion, this comprehensive analysis aimed to assess and compare the performance of various machine learning models for fire risk assessment using a dataset containing diverse features related to environmental factors. The models considered include K-Nearest Neighbors (KNN), XGBoost, Random Forest, Decision Tree, and Support Vector Machines (SVM). Each model underwent rigorous hyperparameter tuning and evaluation through cross-validation to optimize performance.

The KNN model exhibited outstanding accuracy, achieving a perfect score on the test set. However, its simplicity may limit its ability to capture complex relationships in the data. XGBoost demonstrated robust performance, attaining an accuracy of approximately 93.7%, making it a robust choice. Random Forest, while achieving an accuracy of 87.9%, displayed reliable performance, especially in handling diverse datasets.

The Decision Tree model, with an accuracy of 57.8%, highlighted moderate performance, indicating potential limitations in capturing intricate patterns within the data. Lastly, the SVM model performed well with an accuracy of 88.8%, showcasing its adaptability to various kernel types and suitability for both linear and non-linear relationships.

Considering the overall performance, XGBoost emerges as the most promising model for fire risk assessment in this analysis. Its ability to handle complex relationships, robustness in diverse datasets, and high accuracy make it a compelling choice. However, the selection of the most suitable model may depend on specific requirements, computational resources, and interpretability preferences. It is advisable to consider the unique characteristics of the application domain and dataset when making the final model selection.

Add in reference list file and check for Harvard