# Developing Land Use Classification Models Using Machine Learning Techniques: The Case of Asella Town

By: Firaol Feyisa Hamda

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment for Degree of Masters of Science in Computer Science and Engineering

Office of Graduate study

Adama Science and Technology University

Adama

January 2022

# **Developing Land Use Classification Models Using Machine Learning Techniques: The Case of Asella Town**

By: Firaol Feyisa Hamda

Advisor: Dr. Bahiru Shifaw (Ph. D)

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement Degree of Masters of Science in Computer Science and Engineering

Office of Graduate study

Adama Science and Technology University

Adama

January 2022

# APPROVAL SHEET

I, the advisors of the thesis entitled " Developing Land Use Classification Models Using Machine Learning Techniques” and developed by Firaol Feyisa, at this moment certify that the board of examiners' recommendations and suggestions have been correctly integrated into the final version of the thesis.

| Advisor | Signature | Date |
|---|---|---|

We, the undersigned, members of the Board of Examiners of the final open defense by Firaol Feyisa have read and evaluated their thesis entitled" Developing Land Use Classification Models Using Machine Learning Techniques" and examined the candidate. This is, therefore, to confirm that the idea has accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Computer Science and Engineering.

| | | |
|---|---|---|
| Chairperson | Signature | Date |
| Internal Examiner | Signature | Date |
| External Examiner | Signature | Date |

Finally, approval and acceptance of the thesis are contingent upon submitting its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate. Council (DGC) and School Graduate Committee (SGC).

| | | |
|---|---|---|
| Department Head | Signature | Date |
| School Dean | Signature | Date |
| Office of Postgraduate Studies, Dean | Signature | Date |

# DECLARATION

I, hereby declare that this MSC thesis titled "Developing Land Use Classification Models Using Machine Learning Techniques" is my original work. That is, it has not submitted for the award of any academic diploma or certificate in other universities. Furthermore, all sources of materials for the thesis have duly acknowledged through citation.

| | | |
|---|---|---|
| ______________________ | ______________ | ______________ |
| Name of the Student | Signature | Date |

# RECOMMENDATION

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled "Developing Land Use Classification Models Using Machine Learning Techniques" prepared under my guidance by Firaol Feyisa Hamda submitted in partial fulfillment of the requirements for the degree of Mater's of Science in computer science and engineering.

Therefore, I recommend the submission of revised version of the thesis to the department following the applicable procedures.

| | | |
|---|---|---|
| __________________________ | ________________ | _______________ |
| Advisor | Signature | Date |

# ACKNOWLEDGMENT

I am grateful to the almighty for giving me good health and strength through the thesis journey. Furthermore, my sincere gratitude to my advisor Dr. Bahiru Shifaw (Ph. D), for his willingness to advice on the research; his encouragement, guidance, support, and critical comments improved this thesis work.

I want to commemoration the paper to Nebiyou Feyisa Hamda for what he has done for me in life and may Allah bless his soul with paradise, and I would like to thank my wife and son for their support through the journey. Finally, I want to thank my classmates for their support until the completion of the thesis.

Finally, yet importantly, I would like to thank my family for their continuous support and blessing, thanks to all my friends who supported me during the completion of the thesis.

# List of Tables

# List of Figure

# List of Abbreviation

| | |
|---|---|
| ANN | Artificial Neural network |
| CBA | Commercial Business Area |
| CBD | Central Business District |
| CNN | Convolution Neural Network |
| CSA | Central Statistical Agency |
| DL | Deep Learning |
| DT | Decision Tree |
| ELU | Existing Land Use |
| ET | Extra Tree |
| FAO | Food Aid Organization |
| Fuzzy ARTMAP | Fuzzy adaptive resonance theory-supervised predictive mapping |
| IQR | Inter quartile Range |
| LAS | Land Administration System |
| LDA | Linear discriminate analysis |
| LMS | Land Management System |
| LR | Logistic Regression |
| KNN | k-Nearest Neighbor |
| ML | Machine Learning |
| NB | Naive Bayes |
| PLU | Proposed Land Use |
| PSLU | Proposed Structural Land Use |
| RF | Random Forest |
| SLU | Structural Land Use |
| SMOTE | Synthetic Minority Oversampling Technique |
| SVM | Support Vector Machine |
| USGS | United State Geological Survey |
| XGB | Extreme Gradient Boosting |

## Abstract

Land is the most vital resource on earth plays significant role through Economic, Social, Political, Cultural and technical framework dimensions within must good land management and administration. By the 818/2014, urban land holding adjudication and registration need land reform. The reforms include urban land use changing and transformation up to the social interest of the administrative area. The Asella town 31% proposed structural plan contradict with the existing situation. In such case reclassification, proposed structural plan is necessary according to governmental urban plan policy. The roe of land to Asella town to bring sustainable development of the town is crucial.

The study preformed balancing using SMOTE over sampling and implemented stratified kfold for cross validation score. Then the model designed using parametric KNN, RF, DT and SVM and none parametric NB machine learning algorithms used to evaluate the reclassification performance of the designed model. According the result observed from the study KNN, RF and DT score high efficiency more than 99% of accuracy, precision, recall and f1-score. SVM and NB scored lower efficiency compared to other models. In general, the study's findings suggest that machine learning prediction models can identify land use classes and according to the experiment, KNN, RF and DT models are most suitable for land use categorization in respect of Asella town land-use administration undertakings.

**Keywords**: Land use, Land use classes, Cadastre, Classification, reclassification, machine learning classifier

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background of Study

Land is the most vital resource on earth from which humankind derives almost all its basic needs. It plays significant role through Economic, Social, Political, Cultural and technical framework dimensions within must good land management and administration operate (Chekole, 2020). Land is a basis for livelihood; a space for interaction; the source for power and a marker of collective identity. There is a need to manage the wealth of every nation, at least 20% of whose gross domestic product (GDP) can come from land, property and construction.

Asella Town, founded in 1945, is categorized secondary level cities of Oromia and found at the South Eastern of Oromia and the country, at about 175 KM from the capital city of the country Finfinnee and it is the capital city of Arsi Zone. The city is one of the regional cities that are implementing the cadastres under command of regional state located at the main transport route that connects the Southeastern part of the country- to Adama and other town. With latitude and longitude of 7°57′N 39°7′E and an elevation of 2,430 meters, this city is located at a geographic latitude and longitude of 7°57′N 39°7′E. Asella town has a total area of 4623.5 hectares as (office, 2007 -2012 E.C.) reported. In 1984, Asella had a population of 32,954 according to the first national and housing population census (CSA, 2021) in 1987. The second census held in 1994, after a ten-year gap, and the total population was 47,391, up 43.81% from 1984. The population increased to 67,269 people in 2007. According to the CSA (CSA, 2021) data, the population increased by approximately 41.94% from 1994. Only 40% of the total population of the town owns land.

Asella town reestablished as own urban administrative town by the urban local government proclamation of the Oromia National Regional state no 651/2003 in 2004. The objectives of Urban Local Governments as stated under article 7/1 of the proclamation promote self-rule or community governance by encouraging the involvement of residents in the overall activity of the city and facilitate conditions in which residents benefit from the development.

Urban Lands Lease Holding Proclamation No. 721/2011 and re-enactment of urban lease holding proclamation No. 272/2002 give the town to form urban land use and administration for the purpose of urban land tenure and the urban land holding registration proclamation No.818/2014 grant the town to register the information of the lands with land administration systems.

Land administration system is a process for recording, and in some countries guaranteeing, information about the ownership of land. A right is something to which some person or group of persons is entitled. The function of land registration is to provide a safe and certain foundation for the acquisition, enjoyment and disposal of rights in land (Enemark, 2009). Land Management Systems (LMS) are institutional frameworks complicated by the duties they must complete, national cultural, political, and judicial situations, and technological advancements. Because effective land governance and property rights system are fundamental to the overall economic and political development process, countries should set up their LMS based on their understanding of property rights, restrictions, and responsibilities to manage the relationship between Land and people (N.E. Ulger, 2018). Every land administration system should include some form of land registration; cadastres is similar to a land register in that it contains a set of records about land in a multi-purpose role to provide a wide range of land related information, such multi-purpose information records can be built around the land parcel as defined by rights of use.

Thus, much effort invested in order to administer and manage land, of which, cadastral systems are one of these efforts, which developed all over the world. Cadastral system is a formal sub-system of land administration that includes the organizational system (a set of professional actors with responsibilities responsible and accountable to carryout cadastral activities and maintain cadastral information systems), procedures and regulations which altogether ensure that the cadastral system is kept up-to-date (Solomon Dargie Chekole, 2020).

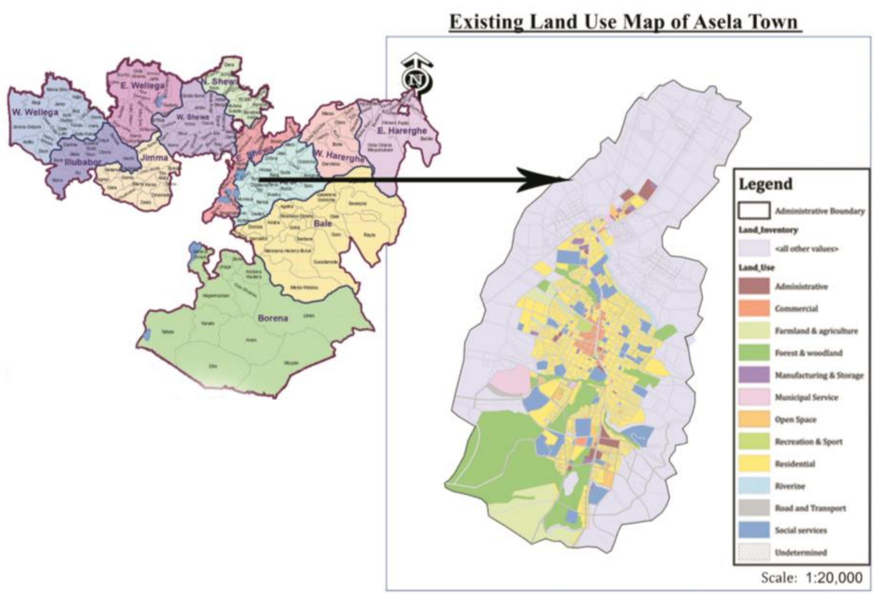
The establishment of urban cadastral system in developing countries is a means of providing security of tenure, by ensuring that people can live with certainty and safety on their own land (Solomon Dargie Chekole, 2020).

According to (Rajabifard, Daniel, & Ian), to achieve sustainable development goals (SDGs) of 2030, countries require access to an effective, efficient and modern land administration system (LAS) based on a cadastre engine that contains spatially accurate land parcels and corresponding rights, restrictions and responsibilities. De Soto argues that the lack of a reliable and efficient cadastral system can have serious implication for the social and economic welfare of a country (De Soto, 2000).

A cadastral system is an information system of land holdings and land use. Therefore, it provides excellent opportunities for identifying problems associated with the development and implementation of land policies. The Cadastre can assist in monitoring and controlling reallocation of land rights to improve social and economic policies through subdivision, land consolidation, land re-allotment, land use change, land transformation, etc. and the value of land as a result of development matters according to (FIG, February 1998).

Ethiopia's urban cadastral system policy is enshrined and incorporated under the urban land development and management policy. This policy a spires a system where urban land is served as a driving force for political, social, economic, and environmental transformation through efficient and well-functioning cadastral system.

To accomplish this vision, the federal government has formulated policies related to urban cadastral systems in order to modernize the system of land administration. Proclamation No. 818/2014 that describes about the urban land holding adjudication and registration ratified to implement urban cadastral system of the town. The formulated policies objective is to accelerate the social-economic and environmental development of urban centers by ensuring landholders' security of holding and recognition of title to immovable property by certifying their right, restriction and responsibility through adjudication and registration. In line with this proclamation, Reg. No. 323/2014 and 324/2014 issued to enact the proclamation. The implementation of urban land holding adjudication and registration need land reform, which is include legal process for implementation of cadastres. Land reform is a form of agrarian reform involving the changing of laws, regulations, or customs regarding land ownership. Land reform may consist of government-initiated or government-backed property redistribution, generally of agricultural land (Land reform). The land reform includes current position regularization, land use changing and transformation to proposed structural land use plans. The regularization performs entitling the parcel with and without certificate with current area of their position according to proposed structural plan.




.

**Figure 1.1 Asella town land use map**

To implement proclamation No. 818/2014, the government has incorporated the urban cadastral system processes in to its GTP to increase urban development.

## 1.2 Motivation

Asella Municipality inefficient land use classification needs a lot of amendment encouraged us to develop a land-use classification model for changing land use classification utilization for the sack of urban cadastres implementation. Therefore, we was inspired to study this problem on the cadastral land inventory data to find a better way to classify the land-use reclassification of the Asella town and recommend for its implementation by the office.

## 1.3 Statement of Problem

In Ethiopia, as in many other African cities, urban growth characterized by a lack of appropriate planning or implementation of plans, resulting in cities with a scarcity of essential infrastructure and services. The lack of urban planning and enforcement, particularly in regional cities, may have contributed to fragmented urban landscape patterns along main highways to their rural edges (Berhanu Keno Terfa, February 8, 2020).

According to our analysis of land use classification of Asella town land use and administration office 31% of parcels, need to amendments as the future land use of the city. These amendments also demo graph the citizen which practical change the geographical covers of the city. To accomplish these amendments and perform land reforms for cadastres purpose land consolidation and reallocation must performed before implementing the town cadastres.

These processes move the landholders to another area or implementing the structural land use to the existing land use patterns. To do so it will need to acquire land and pay compensation either in the form of money or new land given to those who have to move from existing land use to facilitate the development or implementation of cadastres.

The implementation of cadastres clarifies the land use of city must reclassify due to need to acquire land and pay compensation for moved landholder from their current property. According to urban plan proclamation, no 574/2008 of the federal gives the mandate of changing proposed structural land use of specific area, if land use function change found necessary for the public interest, modification of land use function is possible by presenting the plan to the authorized body and upon getting permission under article 21/4. To do so the administration body of city must consider for amending the proposed structural plan as the current land use existing to implement the cadastres process. The amending or reclassification of land uses according to existing society land holding is benefiting the administration by land allocating for the group of new land holders rather than acquire it for compensation and minimize the budget of compensation for implementing proposed structural plan.

This study addresses the problem by answering the following questions:

1. How we prepare the dataset for classification model of land use?
2. What are the most determinant features of land use classification?
3. Which algorithm has better performance for land use classification?

## 1.4 Objectives

### 1.4.1 General Objectives

The general objective of this research is to develop a land-use classification model using machine-learning algorithms.

### 1.4.2 Specific Objectives

The specific objectives are:

- To review the literature on the land use classification and machine learning application in land use classification.
- To collect data related land use from Asella land use and administration offices.
- Preprocess dataset and resample the dataset for tackling the class imbalance.
- Analyze land use data.
- To develop land use classification models and classifying land use by using a different machine learning algorithm.
- Evaluate the performance of developed model.

## 1.5 Significance of the study

Land use is a major contributor to the strain on finite land resources caused by a variety of biophysical and anthropogenic processes, particularly population increase. As a result, the research community, urban planners, stakeholders, and decision-making organizations place a higher emphasis on evaluating and modeling urban land use to have better understand the implications of land use in Asella Town. The findings of this work could help researchers better understand how classify land uses. It also allows you to analyze the classification of land use in the town. Furthermore, the outcomes of this study will serve as a starting point for future research for interested parties in the field.

## 1.6 Scope and Limitation of the study

### 1.6.1 Scope of the study

The study employed on the data manually collected from structural plans and land administration records to classify the town's land use, which decided by the town's future structure plan and the existing land usage practically in the Land. The land use classification classified as “Commercial”, “Industry”, “Public” and “Residential”. The model trained and evaluated using those target classes, and its accuracy, precision, recall, f1-score and confusion matrix used to assess its performance.

### 1.6.2 Limitations of the study

Due to a lack of time, this work will confined to reporting on how to evaluate land use classifications based on accuracy, precision, recall, f1-score and confusion matrix.

## 1.7 Organization of the Thesis

The research paper organized under six chapters in the following ways:

Chapter one discussed above and includes the introduction of the study, the motivation, and the statement of problems, the research questions that would be answered in the proposed solutions, the scope and limitation, the objective of the study, the methodology, and the application results of the study.

Chapter two presents literature review and related works on the definition of Land Use, Land Use Classification, and the factor that determines Land Uses used in the study, Machine learning used in Land Use classification, and finally, it discussed the related work.

Chapter three presents the research methodology used in this research work, methods used to develop Land use classification models which are preprocessing, machine learning classification algorithms, and finally, it presents the evaluation methods.

Chapter four presents the model, preprocessing, balancing dataset, machine learning algorithms and model evaluation implementation for Land use classification solution.

Chapter five presents the visualization results, outliers detection, features correlation, and discussion models evaluation results.

Chapter six presents the conclusion, recommendation and identifies future works for further research direction on this research topic

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 Land Use

The phrase "land use" refers to the human use of Land and the Classification of Land-based on what can be put on it and what can be used for (Land Use, 2022). It depicts the economic and cultural activities (such as agricultural, residential, industrial, mining, and recreational uses) that take place in a specific location as (S. Tadese, February 23, 2021) and (Guoyin Cai, 2019) discussed. Variations in land economic values in metropolitan regions reflect a site's location and geographical advantages, as well as local externalities and governmental rules governing its usage (Nils Kok a, 2014). "The total arrangements of activities and inputs of people conduct in a particular land cover type" according to one definition of land use. It is crucial to distinguish between land use and zoning. Land use refers to how people alter Land to meet their requirements, whereas zoning refers to how the government regulates Land (Duhamel., 2011). Land use restrictions in cities are important determinants of the shape of cities, their physical development and occupancy patterns, people's housing and transportation expenditures, and their economic well-being. Property use restrictions can thus affect land values both directly, through the types of uses allowed, and indirectly, through defining neighborhoods and cities (Nils Kok a, 2014).

Understanding land usage on a greater scale aid in deciphering trends related to Land and urbanization. Understanding how Land has utilized in the past can provide insight into how will it be used in the future. It is safe to assume that humans will always rely on crops and cattle for food, highlighting the need for community land set aside for agricultural purposes (Eastman, August 12, 2020). In urban planning and management, data on land usage plays a critical role. They must update regularly to facilitate decision-making in planning offices at all levels (Qingming Zhan, 2000).

Rapid population growth, as well as the resulting need for housing and other amenities, has increased the usage of urbanized Land, primarily along main roadways and in rural areas, which is often fragmented and inefficient in resource utilization. Infrastructure and service planning can be difficult with dispersed urban development. A compact spatial shape, on the other hand, can aid in the implementation of an improved transportation system as well as reduce traffic energy consumption and resource exploitation, such as water pipe networks and drainage systems. However, some studies claim that compacting urban development has the potential to harm urban green spaces, increase traffic congestion, pollution, and overcrowding services. Even though there is considerable debate over the many types of spatial growth, the consensus is that urban spatial morphology can influence a city's social, environmental, and ecological conditions. As a result, due to its importance in developing sustainable cities, research into urban spatial dynamics and landscape structure analysis continues to be of interest (Berhanu Keno Terfa, February 8, 2020).

According to their application fields and levels, land use classes have been organized hierarchically. Urban built-up areas and non-urban territory will be classed for general use. Residential, commercial, industrial, road, park, and other public green space, among other things, will need to be distinguished for town planning purposes. These land-use classes, of course, have a hierarchical structure. The hierarchical structure can also found in a variety of national standards based on (Nils Kok a, 2014).

The land use classification of Asella town may be analyzed using the benchmark of (Nils Kok a, 2014), however, FAO and USGS have also looked into it. We will talk about it in the upcoming chapter. In addition to natural qualities, the mandatory land use classification distinguishes between social-economic functions. In addition, the town now has seven land use classifications. Residential, commercial, public, open space, greenery, industry, and mixed land uses are the different types of land uses. However, we shall focus on four of them.

Despite attempts dating back to the 1960s, an international agreement on the definition of land use has yet to reach. According to (Jansen, 2006), "the term land use has various meanings among disciplines" and "all of these approaches may be valid." Different definitions can be derived from two techniques, according to (Di Gregorio, 1998): functional and sequential. Land use refers to "the description of the land in terms of socio-economic purpose," according to the function approach. Land use, on the other hand, is defined as "the arrangements, activities, and input humans undertake in a certain land cover type to generate, change, or maintain it" (Jansen, 2006), (Di Gregorio, 1998) under the FAO's sequential approach.

Land use, in any case, is not always easily visible. Land use covers characteristics other than the characterization of the biophysical cover of Land, as defined by the definitions above. Identifying land use involves "socio-economic interpretations of the activities that take place" on the earth's surface, according to (Fisher, 2005). Land use can frequently inferred from basic observations of land cover, but to identify particular Land uses, additional information about human activities on Land or the presence of specific features in the landscape must considered (Jansen, 2006). Land use is distinct from land cover in that some uses are not always visible (e.g., Land used for producing timber but not harvested for many years and forested Land designated as wilderness will both appear as forest-covered, but they have different uses).

Land usage in metropolitan regions is immediately distinguishable as non-rural, implying that agricultural land use is limited. Land usage and function are frequently connected (Nesru H.Koroso, December 2020). The primary land use in almost all metropolitan areas is residential. Industrial land use will be frequent in urban areas, and so forth. However, there are forms unnecessarily related to several urban land uses that the urban area land use classes.

### 2.1.1 Land Use Classification

The majority of land-use classification methods consider both land use and land cover. The United States Geological Survey (USGS) created a major land-use classification system with numerous levels of Classification (James R. Anderson, 1976). Within these levels, the categories grouped in a layered structure. Broad land-use categories such as 'agricultural' or 'urban and built-up' Land are included in the most general or aggregated Classification (level I) (Table 2.1). For regional and other large-scale applications, this level of Classification is widely utilized. There are several more detailed (level II) land-use and land-cover classifications within each level I class. The 'urban and built-up' category, for example, comprises the ‘residential,' 'commercial,' and 'industrial' subcategories. Even more class extensive classes (levels III and IV) can be developed and mapped within each of the level II classes. Each level's classes are mutually exclusive and exhaustive. That is, within each level, each site inside the mapped area can classify into only one class. These four-classification levels are work together to provide a hierarchical system for describing, monitoring and predicting land-use and land-cover change. This multilevel classification approach allows for spatially explicit comparisons of land-use inventories done across time.

Table **2.1** Land cover and land use classification level

| Level I | Level II |
|---|---|
| 1. Urban or built-up | 11. Residential |
| | 12. Commercial and services |
| | 13. Industrial |
| | 14. Transportation, communication, and utilities |
| | 15. Industrial and commercial complexes |
| | 16. Mixed urban and built |
| 2. Agriculture | 21. Cropland and pasture |
| | 22. Orchards, groves, vineyards, nurseries, and ornamental horticultural areas |
| | 23. Confined feeding operations |
| 3. Rangeland | 31. Herbaceous rangeland |
| | 32. Shrub and brush rangeland |
| | 33. Mixed rangeland |
| 4. Forest land | 41. Deciduous Forest land |
| | 42. Evergreen Forest land |
| | 43. Mixed Forest land |
| 5. Water | 51. Streams and canals |
| | 52. Lakes |
| | 53. Reservoirs |
| | 54. Bays and estuaries |
| 6. Wetlands | 61. Forested wetlands |
| | 62. Nonforested wetlands |

| | |
|---|---|
| 7. Barren | 71. Dry salt flats |
| | 72. Beaches |
| | 73. Sandy areas other than beaches |
| | 74. Bare exposed rock |
| | 75. Strip mines, quarries, and gravel pits |

Source: After Anderson JR, Hardy EE, Roach JT, and Witmer E (1976) A Land Use and Land Cover Classification System for Use with Remote Sensor Data. Geological Survey Professional, Paper 964. Washington, DC: U.S. Government Printing Office (James R. Anderson, 1976).

### 2.1.2 Factor influence land uses

Factors influencing land use are physical (soil fertility, soil drainage, slope angle, aspect, scenery, mineral potential, etc.), economic (distance from markets, highway, river, terminal station, stadium, demand for different uses), and social (population size, legislation, government policies) (Briassoulis, 2017).

#### 2.1.2.1 Factor determine change of land uses in the urban area

Urban land expansion modifies habitats, biogeochemistry, hydrology, land cover, and surface energy balance. In most parts of the world, urban land is expanding faster than urban populations. Whereas urban populations expected to almost double from 2.6 billion in 2000 to 5 billion in 2030, urban areas forecast to triple between 2000 and 2030 (Christopher Bren d'Amoura, August 22, 2017). Rapid economic growth and urban have simultaneously promoted the urban land use. The transformation of urban land use has also facilitated by the changes in urban planning and development policies (Wei, August 1993).

Urbanization's contribution to land use change emerges as an important sustainability concern. Here, we demonstrate that projected urban area expansion will take place on some of the world's most productive croplands, in particular in urban regions in Asia and Africa. These urban regions differ from megacities with populations of thousands or more in two important and fundamental ways: administratively, they consist of multiple contiguous entities with discrete governance structures; biophysically, they are a single continuous urban area whose absolute spatial size creates challenges for urban, land, and transport governance. The rate and magnitude of urban land expansion are influenced by many macro factors, including income, economic development, and population growth, as well as a number of local and regional factors such as land use policies, the informal economy, capital flows, and transportation costs.

Urbanization is shown by the increasing percentage of the population in urban areas according to (Cahya, Martini, & Kasikoen, 2018). This resulted in increased demand for housing. Limited land in the urban area push residents looking for an alternative location of his residence to the urban areas. It is accompanied by a process of land conversion from green area into built-up area. In addition, this situation has been experienced land-use change very rapidly from agricultural areas into urban built-up areas.

The implementation of cadastre is an input for modern land administration. While implement the cadastre, the existing land use must interrelated to the proposed urban structural plan. The implementation of the cadastre enforces the administrative organ amending the existing or proposed structure plan as population rate of holding area of land uses. The change of land use or transformation of the land use made on the unit parcel or as the mass of holding area. By considering these, the Asella town administrative bodies have to transform the land uses of area, which are contradicting with the proposed plan.

## 2.2 Machine Learning

Machine learning is a branch of artificial intelligence that focuses primarily on cognitive learning skills. Many additional parts of A.I. aimed at stimulating human function and intelligence (such as problem solving). However, machine learning (ML) is a subset of technology that uses data to replicate human learning (Bao-wen, 2018). If a computer program's performance at tasks in T, as measured by P, increases with experience E, it is said to learn from experience E with relevance to some class of tasks T and performance metric P. The machine learns in terms of specific tasks T, performance P, and experience E, according to Tom Mitchell's definition (Das, 2017). The four types of machine learning algorithms are supervised, unsupervised, semi-supervised, and reinforcement learning; nevertheless, supervised and unsupervised learning are the most well-known (Oladipupo, August 2010).

### 2.2.1 Supervised Machine Learning

Machine learning used to train a function that maps an input to an output based on sample input-output pairs known as supervised learning (Han J, 2011). To infer a function, it uses labeled training data and a set of training examples. When specific goals are established to be achieved from a specific set of inputs (Sarker I.H., 2020), supervised learning is used, i.e., a task-driven technique. The most typical supervised tasks are "classification," which separates data, and "regression," which has an input, output, and requires the user to find a mapping between the two. The goal of Classification is to place the training data into one of the predetermined categories. A series of coaching examples are included in the training data. In regression, the outputs may be real values, while in Classification; the outputs may be class labels. Supervised learning sometimes referred to as categorization since it involves categorizing an input into two or more groups. A supervised learning algorithm examines the training data and generates an inferred function that applied to fresh cases. In the best-case scenario, the algorithm will be able to correctly determine the category labels for unknown cases. This necessitates the training algorithm's ability to "reasonably" generalize from the training data to unknown scenarios. For classifying software instances as defective or defect-free, a variety of classification approaches are commonly used. It is the task of inferring a function from labeled training data, which is performed through machine learning. A collection of coaching examples makes up the training data. Each example in supervised learning could consist of a pair of input objects and desired output values. As (Oladipupo, August 2010) stated both classification and regression are well-known supervised learning tasks. Classification is concerned with the construction of a predictive model for a discrete-range function, whereas regression is concerned with the construction of a continuous-range model. Text categorization, for example, is an example of supervised learning. It involves predicting the class label or mood of a piece of text, such as a tweet or a product review.

#### 2.2.1.1 Classification

The process of classifying ideas and objects involves recognizing, comprehending, and arranging them into predetermined groups or "sub-populations." Machine learning programs use several techniques to classify future datasets using pre-categorized training datasets (Khan, 2010). Classification is the challenge of determining which of a set of categories (sub-populations) a new observation belongs to, using a training set of information that includes observations (or instances) whose category membership is suspected (Khan, 2010). Many real-world problems can be modeled as classification problems, such as classifying a particular email as spam or non-spam, automatically assigning the categories (e.g., "Sports" and "Entertainment") of upcoming news, and diagnosing an illness based on observed features (gender, pressure level, presence or absence of certain symptoms, etc.). Classification is a type of supervised learning in which occurrences assigned to different classes. The output, which in ML is known as a label (Mulugeta, 2015), determines the various classes. In machine learning, classification algorithms use input training data to predict whether the following data will fall into one of the established categories. Filtering emails into "spam" or "non-spam" is one of the most common uses of categorization. In a nutshell, Classification is a type of "pattern recognition," in which classification algorithms are applied to training data to detect the same pattern (similar phrases or attitudes, numerical sequences, and so on) in subsequent sets of data (Kennedy, 2013).

#### 2.2.1.2 Regression Methods

Machine learning is the study of algorithms that can learn from data and make predictions by constructing a model from example inputs rather than following pre-programmed instructions. There are three types of learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. The system is supplied with example inputs and outputs in supervised learning to create a function that maps the inputs to outputs. The two primary types of supervised learning problems are regression and Classification (Bao-wen, 2018).

The unknown parameters in a regression process are frequently termed as, which might be a scalar or a vector. The independent variables are represented by the input vector X, while the dependent variable is represented by the output vector Y. These parameters take the form of vectors when numerous dimensions are involved. A regression model establishes the following approximated relationship between X, β., and

$$\mathbf{Y: Y = f\ (X, \beta)} \qquad \textbf{Equation 1}$$

An anticipated value E (Y | X) is used to approximate the function f (X, β). The knowledge of the relationship between a continuous variable Y and a vector X used to create the regression function f. If no such knowledge exists, an approximation form for f is used.

Consider the k components, also called weights, in the vector of unknown parameters. Depending on the relative magnitude between the number N of observed data points of the form (X, Y) and the dimension k of the sample space, we have three models to relate the inputs to the outputs.

The first instance to model in a regression process is when $N < k$, in which case most traditional regression analysis methods can be used. There is not enough data to recover the unknown parameters β since the defining equation is underdetermined. The equations $Y = f\ (X, \beta)$ can be solved exactly without approximation in the second situation when $N = k$ and the function f is linear because there are N equations to solve N components in β. As long as the X components are linearly independent, the solution is unique. If f is nonlinear, there may be numerous solutions or none at all. The final scenario is when the number of data points (N) exceeds the number of data points (k). This means that the data contains enough information to estimate a unique value in an over determined situation.

### 2.2.2 Unsupervised Machine Learning

The goal of unsupervised machine learning is to find hidden patterns in unlabeled data. In situations where the data categories are unknown, this strategy is applicable. The training data not labeled in this case. Unsupervised learning is a statistic-based learning strategy that addresses the problem of uncovering latent structure in unlabeled data (Oladipupo, August 2010).

### 2.2.3 Semi-supervised Machine Learning

The given data in this sort of learning is a mix of classified and unclassified data. This mixture of labeled and unlabeled data utilized to create a data categorization model that is acceptable. Labeled data is scarce in most situations, but unlabeled data is plentiful as (Brownlee, 2016) stated. The goal of semi-supervised Classification is to create a model that can better predict future data classes than a model created solely from labeled data. We learn in a way that akin to semi-supervised learning (M. Ahmadlou, 2018) and (Oladipupo, 2010).

### 2.2.4 Reinforcement Machine learning

The reinforcement learning strategy tries to adopt behaviors that maximize reward or reduce risk based on observations gained from interactions with the environment (Barto, 2014, 2015). Reinforcement learning goes through the following processes to develop intelligent programs (also known as agents): The agent observes the input state as the initial phase. The decision-making function utilized in the second stage to get the agent to take action. The agent rewarded or reinforced by the environment after completing the activity in the third phase. The state-action pair information regarding the reward then saved as the final step in the (Oladipupo, August 2010) discussion.

### 2.2.5 Machine Learning in Urban Land uses

In recent years, several scholars have used machine learning in urban planning and design to build and monitor city planning. Machine learning utilized for a specific domain in urban analytics and simulations to predictive analysis. Simulators have used to help urban planners make better strategic planning decisions using a variety of methodologies and algorithms. As (Daud, 2021) highlighted the necessity of excellent forecast accuracy in monitoring land-use change in a recent study. The same study also highlighted the requirement for a model that is appropriate for various applications to assure accuracy and fitness.

#### 2.2.5.1 Machine Learning used in Urban Land Use Prediction

In urban planning, the terms "urban development" and "urban extension" used to describe the growth of a specific land area. The examination of land-use change patterns has regained popularity in recent years. Uncontrollable human activities on the land surface were causing environmental damage. As a result, urban simulation employing geospatial data is once again popular. Machine learning now used to improve the accuracy of existing machine learning models and to obtain better classification. The performance of machine learning models determined by the features used to measure land-use change simulation. Predictors are the features employed in the urban simulation model to simulate and predict land usage in this study. The use of these predictors to simulate urban growth patterns allows for an early assessment of urban planning before it is implemented (F. Karimi, 2019). The use of the model in the defined area determines the role and selection of predictors.

Table **2.2** Supervised Machine Learning Classification algorithms

| **Algorithms** | **Advantages** | **Disadvantages** | **Application** |
|---|---|---|---|
| LR | A simple to model, handles a largest number of features and takes little time to training. | Used only for binary classification and poorly responded for multi classification problems. | Credit scoring, predicting user behavior, and discrete choice analysis. |
| KNN | Apply to largest dataset, simple for implement and good for clustering. | Limited on provided initial k value and easily affected by outliers. | Identifying similar documents and detecting outliers. |
| DT | Simple representation of data is easier to interpret and explain to executives, mimic the way humans make decisions in everyday life, smoothly handle qualitative target variables, and handle non-linear data effectively. | Irrelevant to creating complex trees sometimes and do not have the same level of prediction accuracy as compared to others. | Applicable for sentiment analysis and product selection. |

| | | | |
|---|---|---|---|
| RF | Efficient for largest dataset, estimate the significance in classification of input variables and more accurate than DT | Implementations are complex and take more time to evaluate efficiency. | Detect credit card fraud, stock market prediction and product recommendation. |
| SVM | Easy to training the dataset and perform well in high dimensional features | Not perform well when data has noisy elements and sensitive to kernel functions. | Used to detect, identify images and handwritten characters. |
| NB | Require small dataset, perform well in case of categorical input variables and extremely fast compared to methods that are more sophisticated. | Bad estimator | Work well in many real-world situations such as document classification and spam filtering. |
| SGD | Efficiency and ease of implementation and useful when the number of samples is very large | Requires several hyper-parameters and is sensitive to feature scaling | Support different loss functions and penalties for classification. |

## 2.3 Related Works

(Leta, Demissie, & Tränckner, 2021) based their land use classification Digital Elevation Model, Landsat Images, and field data and the dataset sources are based on the three image of different times downloaded from USGS. On the dataset they apply land change modeller. LCM integrated in TerrSet Geospatial Monitoring and Modeling System assimilated with MLP and CA-Markov chain have been used for monitoring, assessment of change, and future projections. The validation accuracy was determined using the Kappa statistics and agreement/disagreement marks. They found average changes of 39.15% in the study area from 1992 up to 2019. They do not use and machine learning algorithm to determine their efficiency.

(Omeiza, 2019) based their land use classification dataset on the NWPU-RESISC45 test dataset, which contains 1050 samples and is 15% of the original dataset using CNN algorithms model. On the test dataset, the model's accuracy, score, and F1 score were

91.03 percent, 99 percent, and 90 percent, respectively. This is an improvement over the chosen benchmark, which used a VGGNet-16-based transfer learning strategy to obtain a test accuracy of 90.36 percent. Their model performance decreased while they try to improve the chosen benchmark.

Based on the classification feature data sets extracted from the multi-source data, combined with field sampling photos and Google Street Views to form various land-use classification data sets, (W. Mao, August 31, 2020) they assess that the RF had the best effect on urban land-use Classification, followed by SVM, and ANN was comparatively poor. With RF the land-use classification training accuracy, validation accuracy, and Kappa coefficient ranged from 91.74 percent to 92.88 percent, 71.89 percent to 79.88 percent, and 0.664 to 0.738, respectively; with SVM, 81.83 percent to 82.34 percent, 68.64 percent to 78.40 percent, and 0.630 to 0.720; and with ANN, 99.60 percent to 99.80 percent, 63.02 percent up to 71.30% percent, 0.559 up to 0.624 from different land use classification level.

(Abdi, 2020) analyze the SVM, RF, XGB, and DL based on the dataset collected for each land-cover class using random sampling from orthophotos of the study area captured on July 2nd and 3rd, 2015 by the Swedish Cadastral and Land Registration Authority (https://zeus.slu.se). The tested algorithms produced similar overall accuracies ranging between 0.733 and 0.758. A third of algorithm pairings were statistically distinct from one another, according to the Z-test comparison of classifier accuracies, whereas McNemar's test results revealed that 62 percent of class wise predictions were significant. Because of the algorithm's modest number of difficult decision boundaries, support vector machines produce the highest classification accuracy.

Decision tree algorithms classification has great potential for land cover and land use mapping based on satellite datasets employing multi-spectral RS-1D/LISS III picture of Heggadadevanakote taluk study area in India, according to (Prasad, May 2012). The Overall Accuracy of the Decision Tree Classified Image was 87.11 percent, with Kappa statistics of 0.8515. The study found that employing remote sensing data, the Decision

Tree might give an accurate and efficient way for classifying land use and land cover mapping.

(Swapan Talukdar, 2020) looked at the overall accuracy (in %) for all classifiers using the Kappa coefficient (K) for the dataset from Landsat 4-5 T.M. and Landsat 8 OLI for the river Ganga from Rajmahal to Farakka barrage in India. Among all the classifiers, the RF classifier was identified as the most accurate LULC model with a Kappa coefficient of 0.89, followed by ANN (K = 0.87), SVM (K = 0.86), fuzzy ARTMAP (K = 0.85), SAM (K = 0.84), and MD (K = 0.82). Between the classified LULC map and ground reality, the R.F., ANN, and SVM models show outstanding agreement, while the other models show very good agreement. Although all of the models are effective, the RF. method is the best LULC classifier.

(Flávio F. Camargo, 5 July 2019) found that as the number of training samples in the dataset on polarimetric ALOS-2/PALSAR-2 L-band pictures obtained on May 14 2016 in the StripMap (SM2) rose. In terms of UA and PA, the RF, MLP, and SVM algorithms did their accuracy. Regardless of the amount of data, the MLP, RF, and SVM classifiers had the greatest Kappa indices (Kappa > 0.50). With 200 samples, the greatest Kappa value for the SVM algorithm was 0.68. When the number of samples is less than 50, the N.B. classifier performs better than the DT J48 classifier, but when the number of samples is greater than 50, it performs similarly or worse. The NB and DT J48 algorithms similarly affected. The NB performed well in the UA With conditional Kappa indices of 0.65 and 0.70 to distinguish between reforestation and natural grasslands. These results were superior to those achieved by other classification algorithms.

### 2.3.1 Gaps of related works

Previous research focused on classification land use based on change happened through time not on current situation of land use transformation to another for implementation of urban cadastre policies. In addition, they based on the information extracted from image using heat map of the area. This study does not show the potential feature determinate of the classification. Most related studies did not collect information on land classification and their data is not stored for the future use.

Table **2.3** Gaps of Relate Literature review

| Author | Title | Model used | Gaps of the studies |
|---|---|---|---|
| (Omeiza, 2019) | "Efficient Machine Learning for Large-Scale Urban Land-Use Forecasting in Sub-Saharan Africa" | CNN | The study does not consider potential determinate. |
| (W. Mao, 2020) | "Comparison of Machine-Learning Methods for Urban Land-Use Mapping in Hangzhou City, China" | RF, SVM and ANN | Small dataset size used |
| (Abdi, 2020) | "Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data" | SVM, RF, XGB and DL | Only 1% of collected dataset is used for the implementation |
| (Prasad, May 2012) | "Decision Tree Classification Model for Land Use and Land Cover Mapping a Case Study" | DT | Spectral (RGB) values of pixel included only in training dataset. |
| (Swapan Talukdar, 2020) | "LULC Classification by Machine Learning Classifiers for Satellite Observations A Review" | ANN, SVM, FA and | Size of the dataset not specified. |
| (Flávio F. Camargo, 5 July 2019) | "A Comparative Assessment of Machine-Learning Techniques for LULC Classification of the Brazilian Tropical Savanna Using ALOS-2/PALSAR-2 Polarimetric Images" | NB, DT, RF, MLP and SVM | The study implemented on very small size of dataset. |

### 2.3.2 Summary literature reviews

The problem identified and the measure taken to solve the problem has been discussed in the above sections' literature review and related works. Many researchers have tried to study land use classification based on aerial image changes using different machine learning.

As discussed in the first chapter, the contribution of this thesis is to study land use reclassification for cadastres implementation in case of Asella town. To overcome this, the proposed work will address the solution using KNN, RF, DT, SVM and NB machine learning algorithm and the gaps will solved by doing this research work.

# CHAPTER THREE

# METHODOLOGY

The strategy for solving the research problem scientifically is known as the research methodology. The procedures used to conduct the study have been described in detail. It includes a review of machine learning methods, a discussion of why classification and multiple regression methods are necessary to predict future land uses, as well as discussions of research design, data sources, data collection techniques, and tools used to design and develop the prototype system. The goal is to make the techniques and steps used to address the research topic clearer. Each technique provides fundamentals that build up to the next, making it easy to grasp how we arrive at our preferred way the methodology we used to develop a land-use classification model using machine learning algorithms.

## 3.1 Research methodology process

The figure 3.1 shows the procedure to achieve the general and specific objectives.

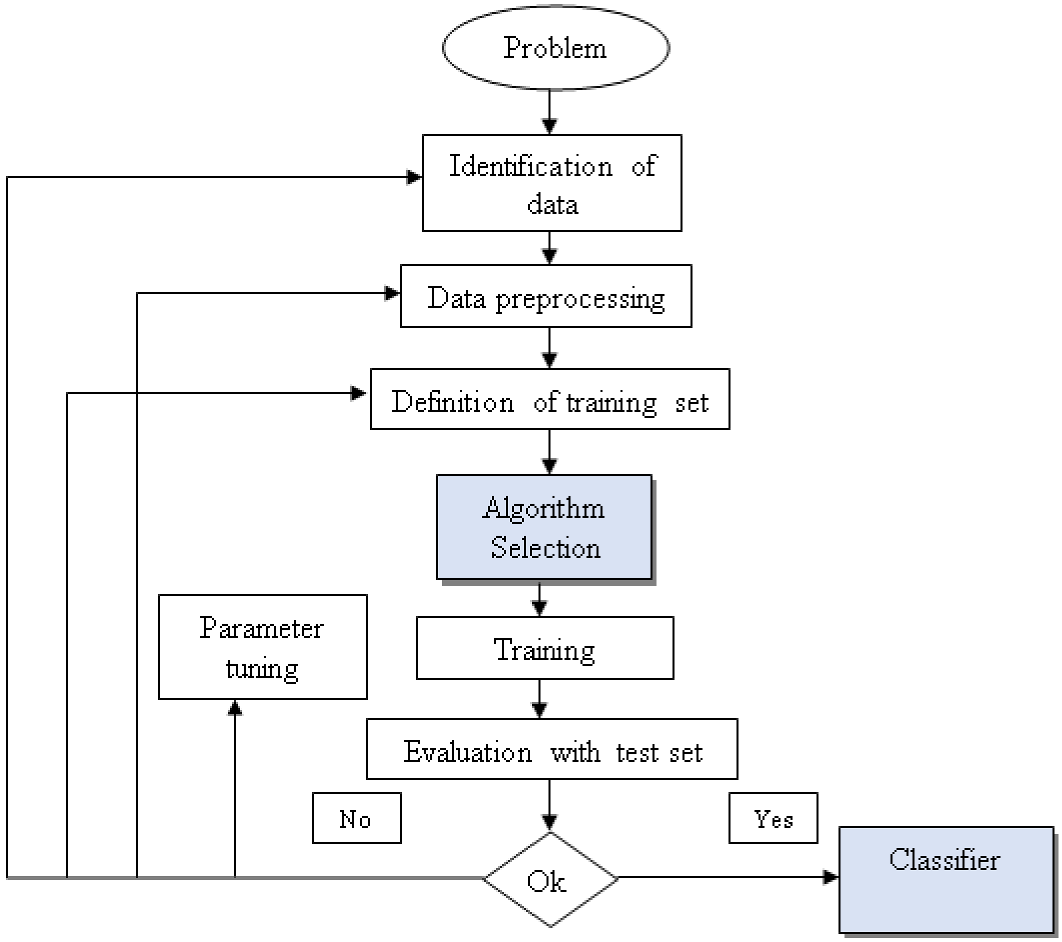


Figure **3.1** Machine Learning Supervise Process (Oladipupo, August 2010)

## 3.2 Tools used

In this research work hardware and software tools used in order to achieve the specific objective of the proposed work for implementing codes, data preparation and drawing diagrams easily and accurately.

### 3.2.1 Hardware tools

- PC core i3
- RAM: 8GB
- Hard Disk: 500GB

### 3.2.2 Software tools

- OS: Window 10 Pro
- Anaconda navigator python version 4.11.0

- Jupyter notebook version 6.4.5
- Python 3.10.1
- Microsoft Excel 2019
- Microsoft Word 2007 and 2016

## 3.3 Data Source

For this study, the collected data is about 21666 from Asella Town land use and administration offices, which used for cadastral land use purposes. From these data 241 duplicated and 9150, missed value in the datasets cleaned. After the duplication and cleaning performed, a size of 12279 datasets of parcels prepared for the model.

## 3.4 Dataset description

After data cleaning, our data set consists of 12279 data records and sixteen attributes (Kebele, longitude, latitude Sex, Age, Parcel Area, Built-up Area, Length from the main road, Length from nearest River, length from the central market, Landholding method, holding year, landholding type, title, PLU and land use with complete information the table describe features of the datasets. The table 3.1 describes the dataset feature of land use.

Table 3.1 Dataset features Description

| **Feature / attribute** | **Description** |
|---|---|
| Kebele | The address of the parcel where the parcel in part of city |
| Longitude | The parcel global specific X-point coordinate in meter unit |
| Latitude | The parcel global specific Y-point coordinate in meter unit |
| Sex | The owners of parcels are male, female of organization or party |
| Age | The owner's age |
| Parcel area | Area of parcel held by the owner in square meter unit |
| Built-up area | The area of building on the parcel in square meter unit |
| Market Distance | The distance between parcel and center market in kilometer unit |
| Main Road Distance | The distance between parcel and main road in the town in kilometer unit |
| River Distance | The distance between parcel and nearest river in kilometer unit |
| Holding Year | The year owner holds the parcel |
| Holding method | The way owner has got the parcel |
| Title ("Carta") | Is the owner has a certificate for his position or not? Boolean data type |
| Holding Type | In what type of landholding s/he hold the parcel e.g., lease or old position |
| Proposed land use | The classification of land use by proposed plan classes |
| Land use | The classification of land used currently used by landholder |

## 3.5 Data Representation

Asella town has eight kebeles; I represent them by alphabet order with numbers from 1 up to 8. In the town, there are three types in gender type female represented by 0, male with 1, and organization with 2. The town has different landholding types like farming, old position, lease, grazing, and so on. In this paper, two of them have been used old positions are represented by 0 and leases by 1. Different land holding methods are there. In this paper, four of them have been used and they have represented Allotment represented by 1, Gift represented by 2, Inherited represented by 3, and purchasing represented by 4. There are two types of landholder in the town have land-using title ("Carta") and have no land using title ("Carta"). Those having Carta are represented by 1 and those which have no Carta are represented by 0. The town has different plans like CBA, greenery, industry, mixed, public, and residential land uses. They have represented from 1 up 6 respectively with their alphabetic order. In addition, the town used four of six future proposed land use; they are commercial, industrial, public, and residential. In addition, they are represented by number orders of 1 up to 4 respectively. I represented the features discussed above as the following table 3.2.

Table 3.2 Data representation

| Represented numbers | Kebele | Sex | Landholding types | Landholding method | Have land using title | Proposed structural land use | Existing land use |
|---|---|---|---|---|---|---|---|
| 0 | - | Female | Lease | - | No | - | - |
| 1 | Arada | Male | Old position | Allotment | Yes | CBA | CBA |
| 2 | Burkitu | Organization | - | Gift | - | Greenery | Industry |
| 3 | Busta | - | - | Inherited | - | Industry | Public |
| 4 | Chilalo | - | - | Purchased | - | Mixed | Residential |
| 5 | Halila | - | - | - | - | Public | - |
| 6 | Hanku | - | - | - | - | Residential | - |
| 7 | Kombolcha | - | - | - | -- | - | - |
| 8 | Welkessa | - | - | - | - | - | - |

### 3.6 Data Preprocessing

Real-world data is frequently incomplete, noisy, and imbalanced, and it is prone to containing unnecessary and redundant data as well as errors. Data preprocessing is a necessary step in preparing data for a machine-learning model. It also improves the model's performance by converting raw data into a comprehensible format. Furthermore, some modeling strategies are quite sensitive to predictors. Cleaning the dataset, managing missing values, transforming data, data reduction and feature extraction, and balancing the biased classes to the proper group of class labels need reclassification prediction were all done using preprocessing approaches. As a result, it is critical to examine and preprocess data before entering the model.

### 3.6.1 Data cleaning

Data in its raw form is unsuitable for analysis. The data cleaned must done first before it used for predictive modeling. The goal of data cleaning is to improve data quality by detecting and eliminating mistakes and inconsistencies. Data quality issues are frequently encountered a result of data input errors, missing information, duplication, redundant data, or other types of erroneous data. To give access to consistent and correct data, the data cleaning process comprises removing missing or duplicate information, filtering unneeded data, and merging multiple data formats.

Data that is not convenient for model training have been preprocessed to avoid the following properties/criteria of data:

- The parcel has a non-missing area.
- The parcel has a non-missing allotment type.
- The parcel has a non-missing allotment year.
- The parcel has a non-missing land-use type.
- The parcel has a non-missing length from the market.
- The parcel has a non-missing length from the center road.
- The parcel has a non-missing length from the river.

The data in the CSV is examine for null values; because the data source contains missing values, each attribute must be checked using the filters, and missing/blank values must be eliminated by filling with mean and mode according to their behavior, which helps to improve the dataset's accuracy. We employ Excel formulas like Conditional Formatting, Remove Duplicates, Extract Number Only, and others to remove erroneous data from a dataset.

#### 3.6.1.1 Dealing with missing data

In real-world datasets, missing data is prevalent, and it has a significant impact on the final analysis result, potentially making the conclusion inaccurate. There are various kinds of missing data. We should be able to figure out why the data is missing. The data sample can still represent the population whether data is absent at random or if the missing tied to a particular predictor but the predictor has no relationship with the outcome. However, if data is absent in a pattern that is relevant to the response, the model will skewed, rendering the analysis unreliable.

Many approaches to dealing with missing data have been developed (Goldstein, March 17, 2009), (Johnson, 2013) and (Kabacoff, 2011), and they can be grouped into two categories. The first and most straightforward option is simply remove the missing data. If the missing data dispersed at random or linked to a predictor with no association to the response, and the dataset is large enough, the elimination of missing data has little impact on the analysis' success. If one of the aforementioned conditions not met, merely eliminating the missing data is ineffective.

The second technique is to use the rest of the data to fill in or impute the missing data. In general, there are two methods. To fill in the missing value, one way simply utilizes the predictor's average. To forecast the missing value, we can use a learning method like Bayes or a decision tree (Enders, 2013). It is worth noting that the imputation introduces more uncertainty.

#### 3.6.1.2 Dealing with outliers

An outlier is a point of observation that differs from the rest of the data. Outliers might derail a model's capacity to analyze data. Outliers, for example, have a significant impact on data scaling and regression fits. However, if the data range not provided, it is difficult to spot outliers. Data visualization may be one of the most effective methods for identifying outliers. We can point out some suspicious observations and examine whether these numbers are scientifically valid by looking at a figure (e.g., positive weight). Only when there are truly valid reasons can outliers be removed.

Outliers can have many causes, such as:

- Measurement or input error.
- Data corruption.
- True outlier observation.

An alternative to removing outliers is to change data to reduce the impact of outliers. The predictor values transformed onto a new sphere by the spatial sign developed by (Serneels, 2006). It reduces the impact of outliers by ensuring that all observations made at the same distance from the center (Johnson, 2013).

Outlier can be of two types:

1. Univariate
2. Multivariate

Univariate outliers can be found when we look at distribution of a single variable. Multi-variate outliers are outliers in an n-dimensional space. In order to find them, you have to look at distributions in multi-dimensions (RP, 2020).

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

This study implemented Z score and IQR for outlier detection.

1. **Interquartile Range Method**

The concept of the Interquartile Range (IQR) used to build the boxplot graphs. IQR is a concept in statistics that used to measure the statistical dispersion and data variability by dividing the dataset into quartiles.

IQR ranking in 25% and 75% in a data set. The values denoted as Q1 and Q3, respectively Quartile regression is an extension of the least squares' method based on the classical conditional mean model.

It is the difference between the third quartile and the first quartile (IQR = Q3 -Q1). Outliers in this case are defined as the observations that are below (Q1 − 1.5 * IQR) or boxplot lower value of variable or above (Q3 + 1.5 * IQR) or boxplot upper value of variable (RP, 2020). The box plot can visually represent it.

$$Q1-1.5 * IQR \leq x \leq Q3+1.5 * IQR \qquad \text{(Equation 2)}$$

Where x is the residual between the estimated value and the true value, Q1 and Q3 are the quarter quartile and third-quarter quartile of the residual, respectively, and IQR is the inter quartile range of the residual. Residuals outside this range are regard as outliers based on quartile and inter quartile range. It follows that the quartile is more resistant to disturbances since up to 25% of the data can be distributed as far as possible without greatly disturbing the quartile (C. Zhang, 2019) (J. Yang, 2019). In the outlier, result true indicates the outlier, which far from other data or it is abnormal whereas false means data set is normal so it used for implementation.

In simple words, any dataset or any set of observations divided into four defined intervals based upon the data values and how they compare to the entire dataset. A quartile is what divides the data into three points and four intervals.

2. **Standard Deviation Method**

Standard deviation is a metric of variance i.e., how much the individual data points are spread out from the mean. In statistics, if a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations

3. **Z-Score method:**

The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of observed or measured.

The intuition behind Z-score is to describe any data point by finding their relationship with the Standard Deviation and Mean of the group of data points. Z-score is finding the distribution of data where mean is 0 and standard deviation is 1 i.e., normal distribution.

A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean μ and also the population standard deviation σ (Z-score).

This technique assumes a Gaussian distribution of the data. The outliers are the data points that are in the tails of the distribution and therefore far from the mean. How far depends on a set threshold zthr for the normalized data points zi calculated with the formula:

$$Z - score = \frac{(X_i - \mu)}{\sigma} \qquad \text{Equation (3)}$$

Where Xi is a data point, 'mean' is the mean of all X and 'standard deviation' the standard deviation of all X.

An outlier is then a normalized data point that has an absolute value greater than Zthr. That is |Z_score| > Zthr, Commonly Zthr used values are 2.5, 3.0 and 3.5. We will be using 3.0

### 3.6.2 Data reduction and Feature Extraction

#### 3.6.2.1 Zero- and Near Zero-Variance Predictors

Some predictors in real-world datasets simply have a single value, which known as a zero-variance predictor. It is useless for predicting variables, and it can cause the model (e.g., linear regression) to crash or the fit to become unstable (Johnson, 2013). As a result, these non-informative predictors eliminated.

Similarly, near-zero-variance predictors with only a few unique values and a very low frequency of recurrence may need to discovered and removed before being included in the model. These near-zero-variance predictors can found by combining two measurements. One is the frequency ratio, which compares the frequency of the most common value to the frequency of the second most common value. The second is a proportion of the specified predictors' unique values. If the predictor's frequency ratio exceeds a predetermined threshold and the fraction of unique values is low, it can classify as a near-zero-variance predictor.

#### 3.6.2.2 Identifying Highly Correlated Predictors

Co-linearity is a term used to describe a situation in which a pair of predictor variables are highly connected in real-world data (Williams, 2010). For the following reasons, it is critical to pre-process data to avoid strongly correlated predictor pairs in the data. For starters, two highly correlated predictors are likely to contain the same final information, and adding more predictors increases the model's complexity. Fewer predictors are desirable in models were training the predictors is expensive (such as Support Vector Machine and K-Nearest Neighbors). Furthermore, using highly correlated predictors pairs in a model might result in a very unstable model, causing numeric mistakes and poor prediction performance in some machine learning methods.

When dealing with correlated predictors, removing a small number of predictors with the highest pair wise correlations and ensuring that all pair wise correlations are below a given threshold is a good strategy. The main idea is iteratively remove highly correlated predictors, as seen in the algorithm below (Johnson, 2013):

1. Determine the correlation matrix of the predictors.
2. Find the pair of predictors (A and B) with the greatest absolute pair wise correlation.
3. Determine the average correlation between predictor A and the other predictors, as well as the average correlation between predictor B and the other predictors.
4. Discard the one (A or B) having the highest average correlation with the other predictors.
5. Continue Steps 2–4 until all absolute correlations are less than the threshold.

### 3.6.3 Data Transformation and Normalization

#### 3.6.3.1 Data Transformation

Many statistical models assume that the data analyze is normally distributed; however, data might be positively or negatively skewed. Data normalcy improved using transformations to resolve skewness (Osborne, 2002).

There are a variety of transformation approaches that used to correct the skew, including substituting data with log, square root, or inverse transformations (Johnson, 2013). Box and Cox (1964) suggest a set of transformations that can experimentally discover an acceptable transformation by indexing them by a parameter $\lambda$.

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda}, if\ \lambda \neq 0 \\ \log x\,,\quad if\ \lambda = 0 \end{cases}$$ **(Equation 4)**

Equation (1) can identify a variety of transformations, including log transformations, square root transformations ($\lambda$ = 0.5), square transformations ($\lambda$ = 2), inverse transformations ($\lambda$ = -1), and other intermediate transformations.

#### 3.6.3.2 Data Normalization

Normalization used to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values (S. García, 2016). This study implemented StandardScaler normalization.

## 3.7 Machine Learning Classification Algorithms Models

The study's goal is to apply machine-learning algorithms to forecast urban land use patterns. Multiple supervised machine learning algorithms utilized to achieve the goal, and their performance compared. The Classification model built using the following machine learning algorithms. The algorithms were chosen based on their past work results and their strong performance and popularity in land use prediction challenges.

**Decision Tree:** Decision trees frequently graphically depicted as a hierarchical structure, which makes them easier to understand than alternative strategies. This structure primarily consists of a starting node and a series of branches that lead to other nodes until we reach the leaf node, which contains the route's final choice. Because of its straightforward form, the decision tree is a self-explanatory model. Each internal node evaluates an attribute, whereas each branch represents the value of that attribute. At the end of the process, each lead assigns a categorization. A decision tree is a classification algorithm that expressed as a recursive partition (splits) of the input space depending on attribute values. Each leaf allocated to one of several classes, each of which indicates the most appropriate or common goal value (Abdul Fattah Mashat, 2012).

**Random Forest:** The random forest classifier is a meta-estimator that fits several decision trees on different sub-samples of datasets and utilizes average to improve the model's predictive accuracy and control over-fitting. The size of the sub-sample is always the same as the size of the original input sample, but the samples generated with replacement. In most circumstances, a reduction in over-fitting and a random forest classifier are more accurate than decision trees. Real-time prediction is slow, and the technique is difficult to implement (Livingston, 2005) and (Oladipupo, August 2010) indicated.

**Support Vector Machine:** Support Vector Machine algorithm capable of performing classification, regression, and even outlier detection. The training data represented as points in space split into categories by a clear gap as broad as possible in a support vector machine. New examples mapped into the same space and classified according to which side of the gap they fall on. It works well in high-dimensional areas and only utilizes a small number of training points in the decision function, making it memory-friendly. The approach does not offer probability estimates directly; instead, these are derived using a time-consuming five-fold cross-validation as (Nikola Kranjˇci´c, 18 March 2019) and (Lewes, January 2015) stated.

**Naïve Bayes:** The Naive Bayes algorithm assumes that every pair of features is independent based on Bayes' theorem. Many real-world situations, such as document classification and spam filtering, benefit from Naive Bayes classifiers. To estimate the required parameters, this technique requires a limited amount of training data. When compared to more advanced algorithms, Naive Bayes classifiers are extremely fast. Naive Bayes has a reputation for being a poor estimator (Garg, June 2013).

**K-Nearest Neighbors:** Neighbors-based categorization is a sort of lazy learning because it does not try to build a general internal model and instead just saves instances of the training data. A simple majority vote of each point's k nearest neighbors used to classify it. This technique is straightforward to use, robust to noisy training data, and successful when dealing with massive amounts of data. It is necessary to calculate the value of K, and the computation cost is considerable because each instance distanced from all of the training examples (SSCS, 2010) and (Fischer, 1970).

### 3.7.1 Proposed solution architecture model for land use reclassification

We develop the following architecture to classify the land use of Asella town base on the above data processing.  The criteria dealing with missing data and outliers, scaling and normalization, transformation to resolve skewness, zero and near zero variance, and identifying highly correlated predictors performed. We implant SMOTE resample strategies to remove the biasness of model and the stratified cross validation use to train and test split for continuous balancing of our representing classes.

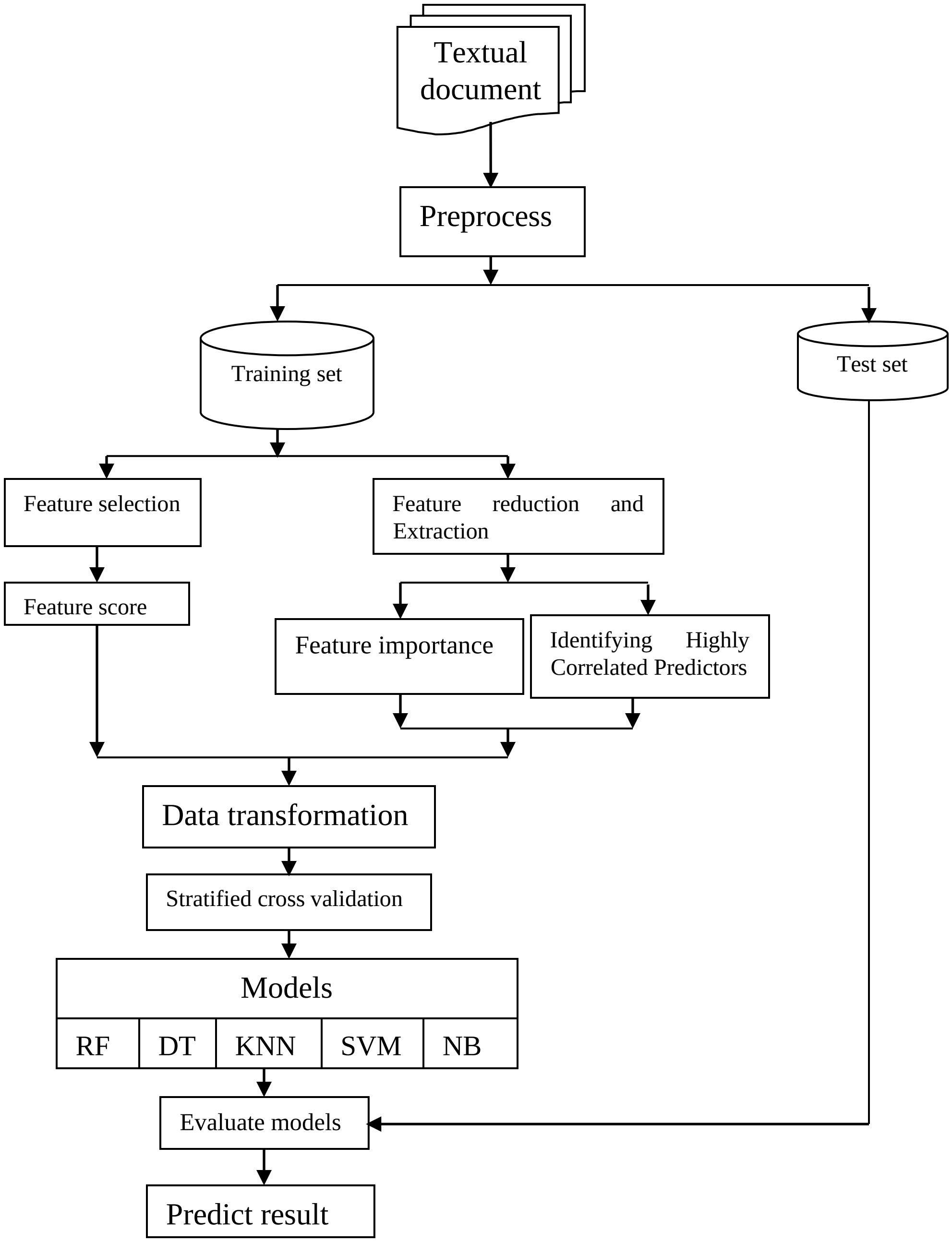


Figure **3.2** Land use reclassification proposed architecture

## 3.8 Model Evaluation

Model evaluation is a critical component that helps us select the appropriate approach for our data. It assists us in determining how well our algorithm will function in the future. This study uses a variety of evaluation criteria to assess the algorithms' performance, including accuracy, precision, recall, f-measure, and confusion matrix. The following words are used to describe these metrics:

- True Positive (TP): when the actual observation is correct but the prediction is correct.
- False Negative (FN): when the observed value is positive while the expected value is negative.
- True Negative (TN): when the observation is negative yet the outcome is expected to be negative.
- False Positive (FP): when a negative observation projected to be positive.

### 3.8.1 Accuracy

The value of correct classification predictions divided by all predictions is called accuracy (ML | Evaluation Metrics, 2021). The formula is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{.................. } \textbf{(Equation 5)}$$

### 3.8.2 Precision

Precision tells you how many of the responses are right and relevant. It displays the number of times the model correctly predicts the outcome (ML | Evaluation Metrics, 2021). It calculated using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} \text{ ............................ } \textbf{(Equation 6)}$$

### 3.8.3 Recall

The number of genuinely relevant proportions is shown in recall (ML | Evaluation Metrics, 2021). It calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN} \text{ ................................ } \textbf{(Equation 7)}$$

### 3.8.4 F-Measure

The equation calculates the F-measure, which is defined as the harmonic mean of precision and recall (ML | Evaluation Metrics, 2021).

**F1-Score: (2 x Precision x Recall) / (Precision + Recall)**

In all types of classification algorithms, the F1-Score is the weighted average of Precision and Recall. As a result, this score considers both false positives and false negatives. F1-score is frequently more valuable than accuracy, particularly if your class distribution is unequal.

- Precision: How often does a positive value prediction turn out to be correct?
- Recall: How often is the forecast true when the actual value is positive?
- $\mathrm{F} - \mathrm{measure} = 2 * \frac{(\mathrm{Recall} * \mathrm{Precisoin})}{(\mathrm{Recall} + \mathrm{Precision})}$ ……… **(Equation 8)**

### 3.8.5 Confusion Matrix

A confusion matrix is a table that displays the classification model's prediction results. By comparing the actual values with the predicted values by the classifier, it summarizes the value of correct and incorrect predictions. It also displays the errors, the types of errors, and how the classifier becomes confused during prediction (ML | Evaluation Metrics, 2021).

**Accuracy: (True Positive + True Negative) / Total Population**

- Accuracy defined as the proportion of accurately predicted observations to all observations. The most intuitive performance metric is accuracy.
- True Positive: The number of right forecasts that the event will occur positively.
- True Negative: The correct predictions number that an event will occur negatively.

## 3.9 Sampling Techniques

We used a probability sampling strategy for this study. The test error used to assess the model's performance. During the fitting process, a subset of the training data held out and then the fitted model applied to the held-out subset. To test and train our dataset, we use 10 stratified k-Folds Cross-Validation sampling technique.

To separate the dataset and each utility, we utilize a sampling strategy and split our dataset with 80% training and 20% testing for modeling.

# CHAPTER FOUR

## IMPLEMENTATION

This research takes a unique experimental approach, utilizing the most advanced machine learning algorithms. Performs seven experiments with two datasets and seven machine-learning algorithms compares the results and recommends the best algorithm.

### 4.1 Model Implementation

We used Python sci-kit learning modules used for training and testing models, Matplotlib.pyplot and Seaborn python modules utilized in this study to visualize the dataset result. We used Imblearn package for data balancing. A multidimensional array, the matrix data structure is included in numerical Python (NumPy) package. To convert data to a NumPy array, use this function. In addition, Pandas allows us to import data from a variety of file formats, including CSV, merge data, and reshape it. After importing every package then we loaded land uses a dataset.

```
df = pd.read_csv('landuse.csv')
```

Figure **4.1** Python code for loading the datasets

### 4.2 Pre-processing implementation

Preprocessing method have been implemented for cleaning the dataset, handling missing values, handling outliers, scaling, normalization, transformation, data reduction and feature extraction for prediction model. In this study, the python programming language used to implement land use prediction model. To perform the implementation, we imported library packages and loaded the dataset used in the implementation by using the following python code.

### 4.2.1 Implementation steps to prepare dataset for the model algorithms.

To clean the missed values in the dataset, first .info() method used to observe the dimensional and row information of datasets. These method return total number of none missed values of the feature. Then .describe() method used to select the way we fill the missed values. These method returns mean, standard deviation, minimum and maximum variance of the features. We can select one of them based on the behavior of features. We replaced missed values of longitude, latitude, parcel area, built up area, market distance, main road distance and river distance with mean value. In addition, the missed values of kebele, sex, age, holding method, holding year, holding types, title and land use values replaced with mode values. After the missed value filled, we applied scaling and normalization to the features. Then .var() method is used to check the variance of features. These method returns the variance of features. We used .log() use for power transformation for highest variance degree of the features and normal scale for lower variance of the features is used. After we scaled and transformed we check zero and near zero variance. In addition, longitude and latitude features removed. Then .corr() method is used to check the correlation of the features. These methods returned the correlation of the features. To remove highly correlated features we applied the .corr().abs() method to find highly correlated features. These methods returned holding year feature.

### 4.2.2 Dataset outliers detection results

We implemented two different outlier detection techniques. They are Interquartile Range (IQR) and Z-scores.

#### 4.2.2.1 Interquartile Range (IQR)

We implemented IQR outlier detection techniques for longitude and latitude dataset variables using boxplot graph and implementing IQR method.

#### 4.2.2.2 Z-scores

We implemented Z-score outlier detection techniques for market, main road and river distance variables using distplot graph and implementing Z-score method.

### 4.2.3 Declaring features and target

Declaring features and targeting the main point of implementation to identify the independent and dependent features. The information can be a two-dimensional numerical array or matrix, which we will call the *features matrix*. By convention, this features matrix is often stored in a variable named x. The features matrix is assumed to be two-dimensional, with [12279 samples, 11 features], and is most often contained in a NumPy array or a Pandas DataFrame, though some Scikit-Learn models also accept SciPy sparse matrices. In addition, we also generally work with a label or target array, which by convention we will usually call y. The target array is usually one dimensional, with length 12279 samples, and is generally contained in a NumPy array or Panda's columns. The target arrays have continuous numerical classes/labels. We will be working with the common case of a one-dimensional target array.

```
#Declare feature vector and target variable
x = landuse_df.drop(['land_use'], axis=1)
y = landuse_df['land_use']
```

Figure **4.2** python code for declaring feature vector and the target variable

## 4.3 Balancing Dataset Implementation

To balance the classes of land uses of selected datasets using the proposed method, the study used imblearn python package that gives access to resampling techniques. SMOTE() sub-module of over_sampling package is used to implement the over-sampling.

SMOTE () class over-sample minority class which are Commercial, Industry and Public land use class by randomly picking samples from the dataset the ratio of over-sampling techniques; in this study, we used equal ratios to SMOTE() over-samples the minority by randomly duplicating the minority class. This means the minority class oversampled as the majority class and the .fit_resample() function is called to fit and apply the transformed data to the dataset.

```python
#resampling with SMOTE
from collections import Counter
from imblearn.over_sampling import SMOTE
smote = SMOTE()
# fit predictor and target variable
x_smote, y_smote = smote.fit_resample(x, y)
```

**Figure 4.3** python code for resampling with SMOTE

After the dataset with land use label is loaded, and we StratifiedKFold using the cross_val_score() method from sklearn.model_selection package. We split the dataset into 10 k-Fold Cross-Validation for continuous equaling of distributed class labels to test the model.

```python
skf=StratifiedKFold(n_splits=10, shuffle=True, random_state=0)
```

Figure **4.4** python code for splitting dataset to train and test set using stratifiedkFold

After the dataset split the stratified kFold used for continuous class label and the model have standardized before further implementation. One way of standardizing the model is make_pipeline from sklearn.pipeline package and StandardScaler() method use for scaling the model. This method transforms a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. StandardScaler results in a distribution with mean as zero and standard deviation equal to one. In short, it standardizes the data. Standardization is useful for data that has negative values. It arranges the data in a standard normal distribution. StandardScaler() method used for feature scaling which is called standardization, this method scales the values for each numerical feature in the dataset.

```python
# Create an instance of Pipeline
scaled_model = make_pipeline(StandardScaler(), model)
```

Figure **4.5** Python code for model standardization

## 4.4 Machine Learning Algorithms Implementation

To implement the RF classifier, RandomForestClassifier() from sklearn.ensemble package is used to train the classifier. Grid search methods are used for hyperparameter tuning, GridSearchCV() from sklearn model selection used to generate candidates from a grid of parameter values specified with the RF param, while fitting the best combination between parameters are evaluated and will be retained. This classifier can fit the training dataset to predict the test dataset.

To implement DT classifier, DecisionTreeClassifier() from sklearn.tree package is used to train the classifier. Grid search methods are used for hyperparameter tuning, GridSearchCV() from sklearn model selection used to generate candidates from a grid of parameter values specified with the DT param, while fitting the best combination between parameters are evaluated and will be retained. This classifier can fit the training dataset to predict the test dataset.

To implement the KNN model we used KNeighborsClassifier() from sklearn. neighbors package to train the classifier. Grid search methods are used for hyperparameter tuning, GridSearchCV() from sklearn model selection used to generate candidates from a grid of parameter values specified with the kNN param, while fitting the best combination between parameters are evaluated and will be retained. This classifier can fit the training dataset to predict the test dataset.

The study used the LinearSVC() classifier of the sklearn.svm package to build the SVM model. This classifier instantiate OVR parameter for multi-class Classification can fit the training dataset to predict the test dataset.

To implement Naïve Bayes model we used GaussianNB from sklearn.naive_bayes package trains the classifier. We do not apply grid search for GaussianNB() classifier because it known as bad estimators. This classifier can fit the training dataset to predict the test dataset with parameter.

## 4.5 Model Testing and Evaluation

The performance of the model evaluated using a stratified kFold implementation on the dataset. The study used an accuracy_score(), confusion_matrix() and classification_report() methods from the sklearn.metrics package are used to evaluate the performance of the built models using predicting the value of the test set. These methods give a result of evaluation methods such as accuracy, precision, recall, and F1-score value of the model under testing.

# CHAPTER FIVE

# RESULT, EVALUATION, AND DISCUSSION

This chapter presents the result obtained from the experiment of land use prediction using an implementation of the supervised classification machine learning algorithms: RF, KNN, DT, SVM and NB. Here the result of dataset balancing techniques and the result of classification models for land use classification will discussed. In the final section, we discuss the result obtained by each experiment in this study.

## 5.1 Result of dataset Visualization

### 5.1.1 The feature distribution of the dataset

The following fig 5.1 result shows the features histogram plot that shows the frequency of features and the variance of datasets. As we see from the figure the variance of longitude, latitude, parcel area, and built-up area have large variance results to transform.

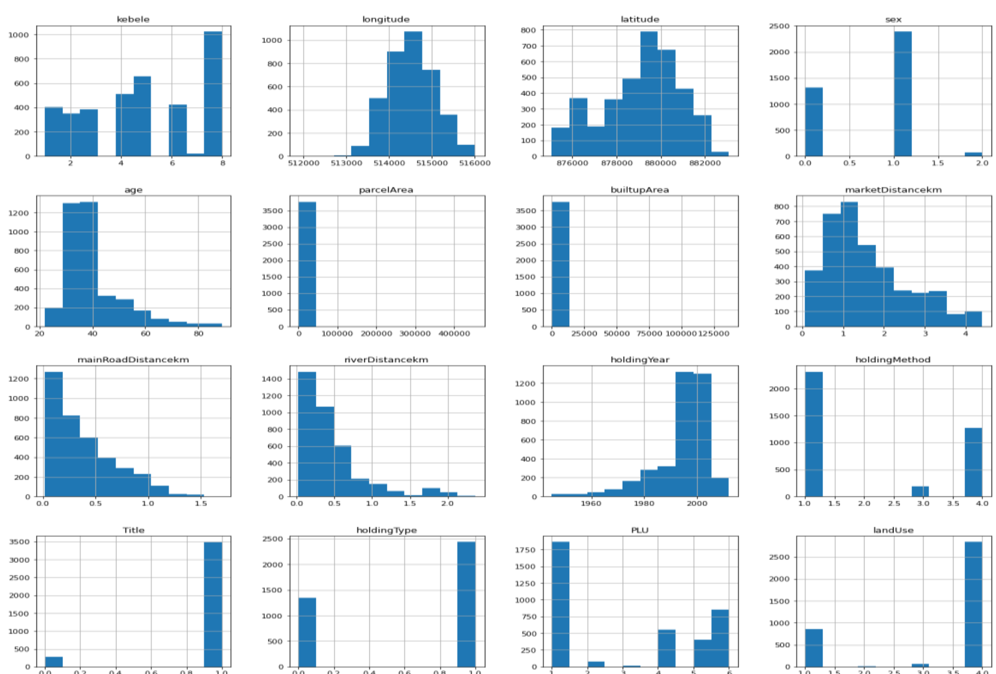



Figure **5.1** the histogram result of features

To handle missed values data in the dataset. We used the count number returned of missing values in the dataset that we observe from fig 5.2. A simple way to deal with data containing missing values or not, in this case the missed value is few, is to fill with missing values in the dataset.

```
kebele                0
longitude             4
latitude              4
sex                   1
age                   3
parcelArea            1
builtupArea           5
marketDistancekm      5
mainRoadDistancekm    2
riverDistancekm       4
holdingYear           0
holdingMethod         3
Title                 4
holdingType           5
PLU                   8
landUse               7
```

Figure **5.2** output result of missed features values.

A Figure 5.3 is a graphical representation of the run command result of features that are different with the numbers of their class's kebele 8 has 2335 counts, kebele 2 has 2158, kebele 1 has 1752 counts, kebele 5 has 1725, kebele 7 has 1468 counts, kebele 3 has 1081, kebele 9 has 909 counts and kebele 4 has 847. The kebele feature 2335 maximum and 847 minimum numbers**.** The longitude feature has an average of 514415.5525. The latitude feature has an average of 878893.2056. The male sex features have 7825 counts the female sex feature and 85 none gander or organization of 85 counts. The average age is 39.24. The average area of parcels is 389.79 square meters. The average built-up area of the parcel is 214.64 square meters. The center of the market has an average of the 1.60-kilometer distance from the parcels. The main road distance has an average 0.58-kilometer distance from the parcels. The nearest river has an average of the 0.47-kilometer distance from the parcels. The holding year of the parcel is between 1945 and 2012. The holding method has a mode of one. The title has a mode of one. The holding type has a mode of one. In addition, the land use has a mode of four. In addition, it fills the missed values with mean and mode. The longitude, latitude, parcel area, built-up area, market, main road, and river distance missed value replaced with mean. The sex, age, holding method, title, holding type and land use missed value replaced with a mode.

To observe the detailed attributes in the feature distribution in the left of figure 5.3 and dataset positively skewed with right of figure 5.3 below. The result shows the distribution and frequency of dataset variable features. This distribution and frequency must treated with outlier detection and normalize.

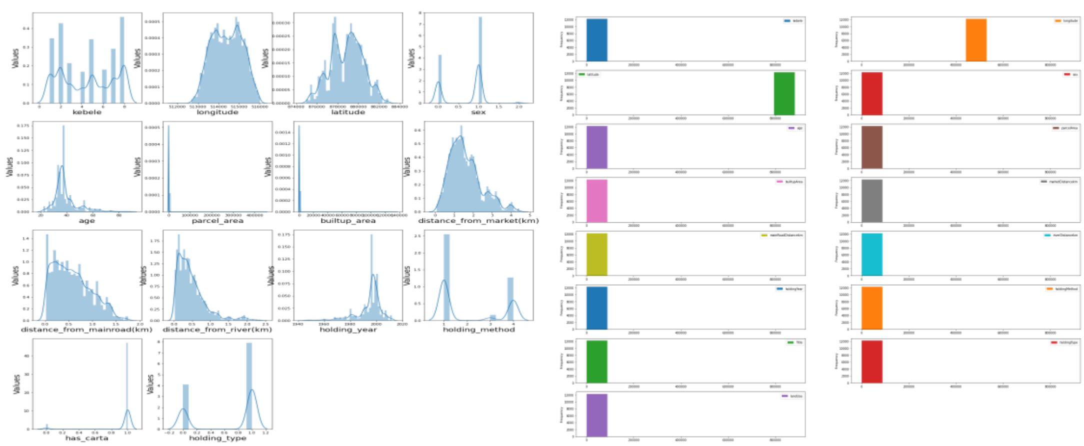



Figure **5.3** The visual display distribution and skewed of dataset features

By visualizing the dataset variables distribution and frequency, we observe normal distribution of longitude and latitude feature variables as below figure 5.4. Then we applied the IQR for longitude feature variables observe the outlier. The latitude feature was free from outlier.

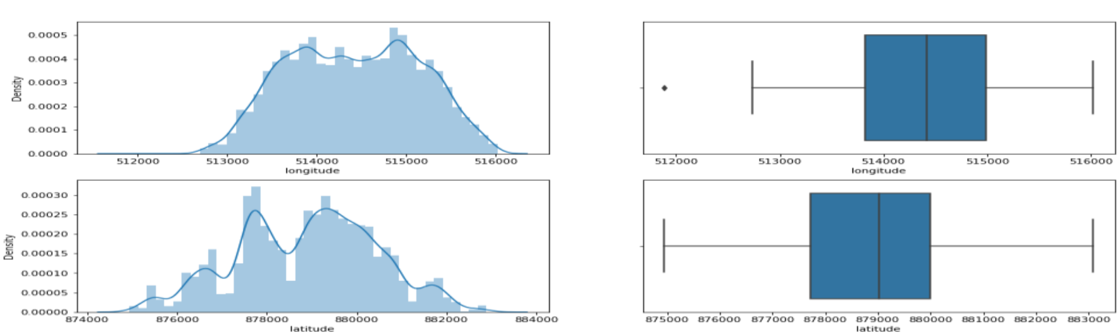



Figure 5.4 the distplot and boxplot result of longitude and latitude features

After we applied the IQR for longitude feature, we removed the outlier and the figure 5.5 show the result of removed outlier from longitude feature variable.

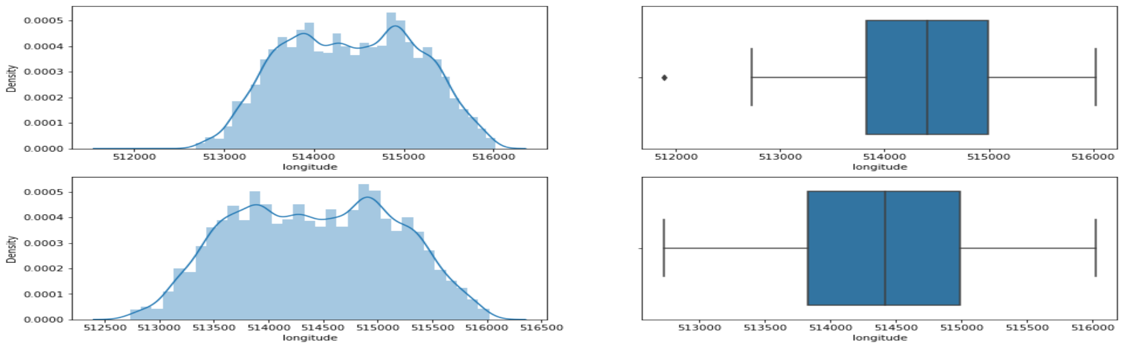



Figure 5.5 the distplot and boxplot result of longitude feature with and without outlier

As we observe the variable distribution of feature on figure 5.3 market, main road and river distance have subnormal distribution of variables. Then we applied Z-score on these features. The figure 5.6 show the distplot result of feature variables of market, main road and river distance.

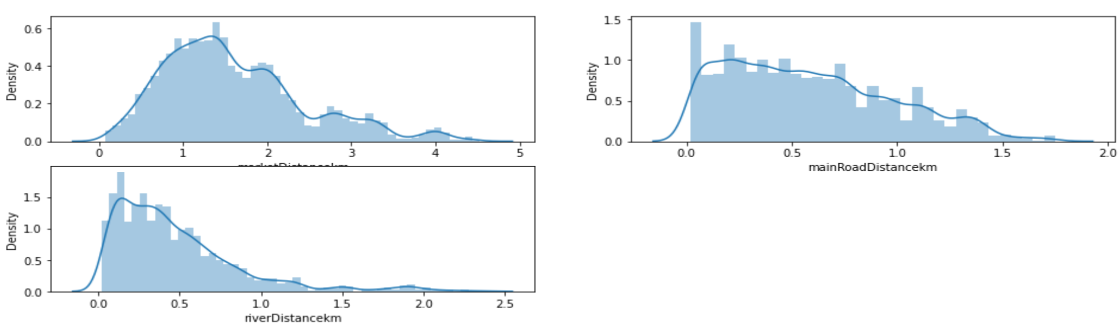



Figure 5.6 the distplot of market, main road and river distance features

Then by applying the Z-score, we removed the standard deviation threshold less and greater than 3.0 mean values of market, main road and river distance variable feature as the figure 5.7 results.

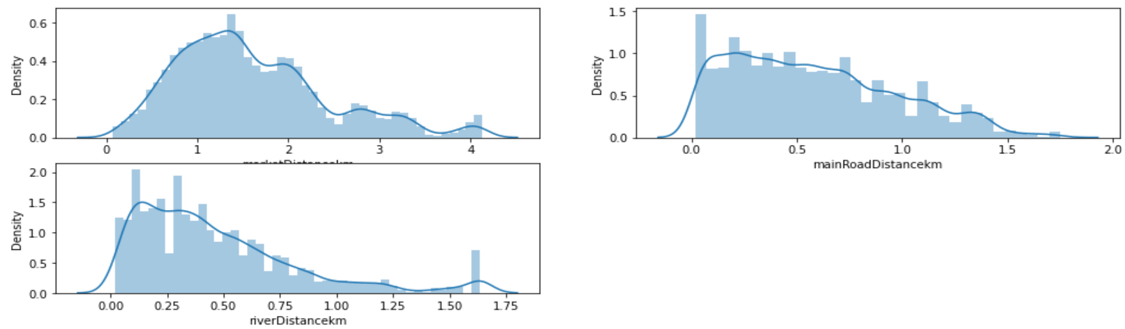



Figure 5.7 the distplot threshold removed result of market, main road and river distance

After we removed the outlier with IQR and Z-score, we applied the mean ratio to parcel and built up area variable features. Then we found the result of parcel and built up area as figure 5.8.

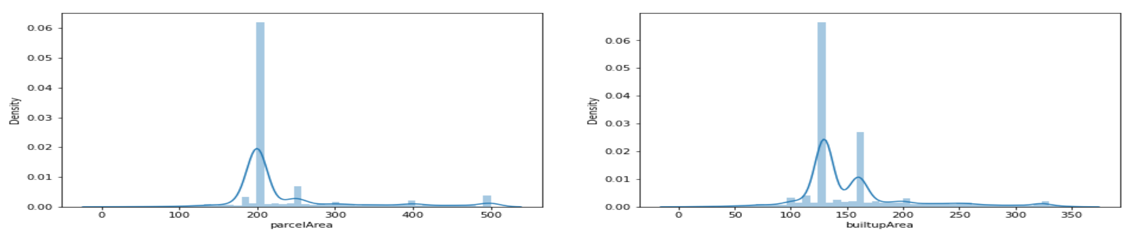



Figure 5.8 the mean ratio distplot result of parcel and built up area.

After we implemented the mean ration to parcel and built up area, then the correlation coefficient or Pearson correlation coefficient is checked a numerical index of the strength of the relationship between variables. Seventeen variables tested in python to check whether this assumption is implementation. One variable is created to check whether the land uses need amendment or not. The fix name given to newly created variable. Fig 5.9 shows the 2D correlation coefficient heat map matrix of removed highly correlated features.

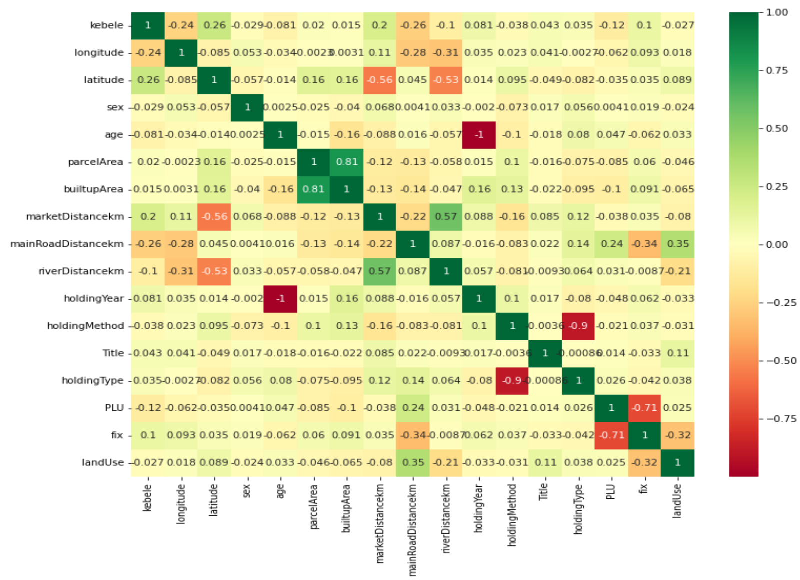


Figure **5.9** the correlation matrix result of the preprocessed of dataset.

The result shows the correlation strength and direction of the relationship between the variables. The correlation coefficient ranges between -1 to +1, thus -1 indicates a negative correlation, and +1 indicates a positive correlation.

As (Williams, 2010) stated the result of correlation analysis is based on the following interpretation: From 0.1 to 0.39 is weak,0.40 to 0.60 moderate, 0.70 to 0.89 strong, and from 0.90 and above is very strong.

According to this guideline parcel area and the built-up area has a strong correlation whereas the market and river distance has a moderate correlation. Kebele, market distance, main road distance and river distance have a weak correlation. The highly correlated features decrease the performance; they removed before implementing the model. The following fig 5.10 shows the correlation scatter plot of features with target.

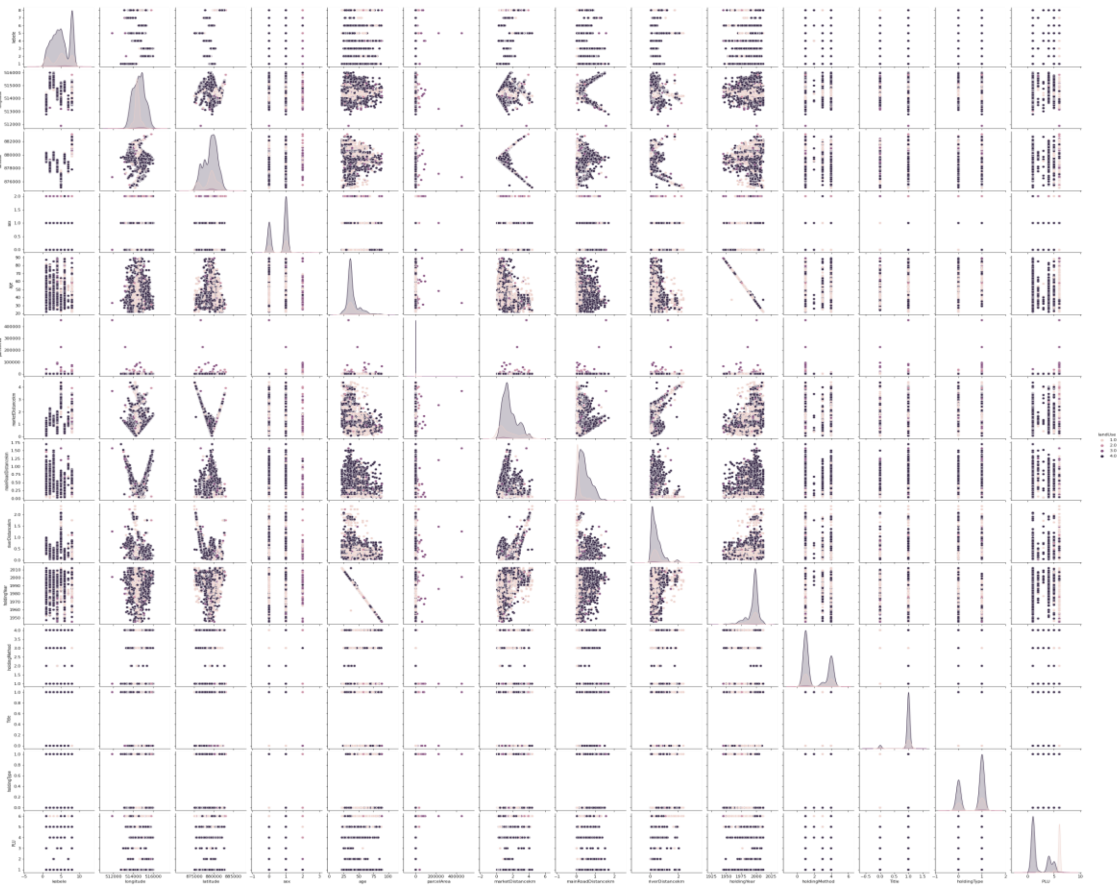


Figure **5.10** the scatter plot of features with the target features

After we removed correlated features of dataset, we check the variance of features. The sex, holding year, holding method, PLU and title features removed because they are nearly zero variance and have not mutual relation with the target features. The feature selection is the processes of identifying features determine the target features. The mutual information of the features is one of the feature selection processes. The Figure 5.11 is the mutual information of features in the dataset.

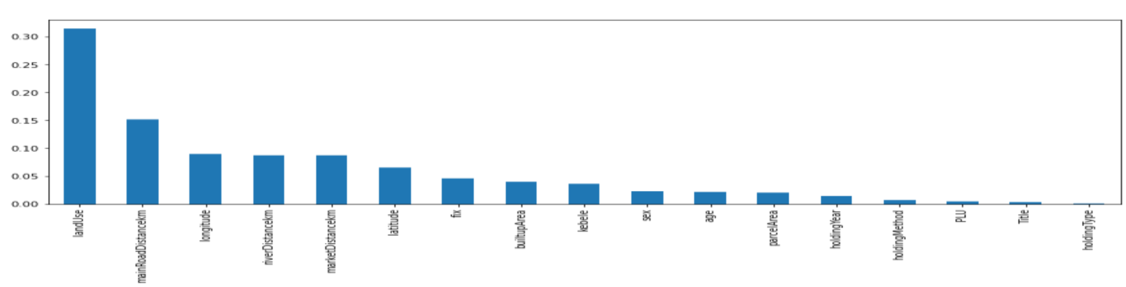
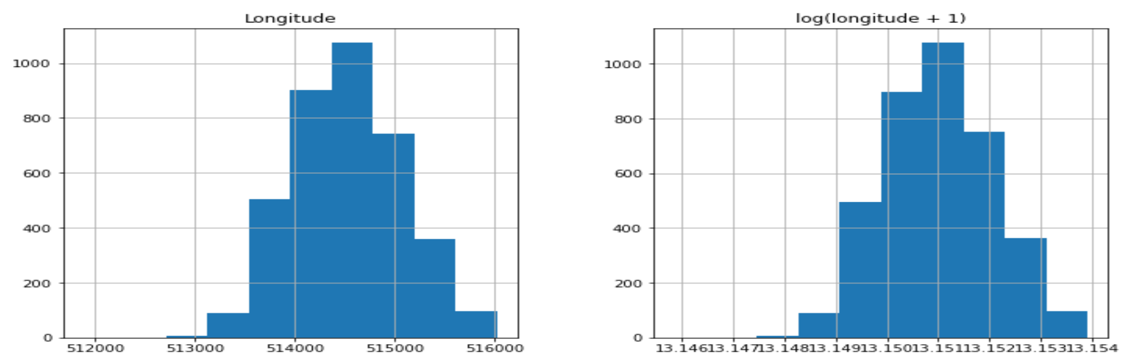
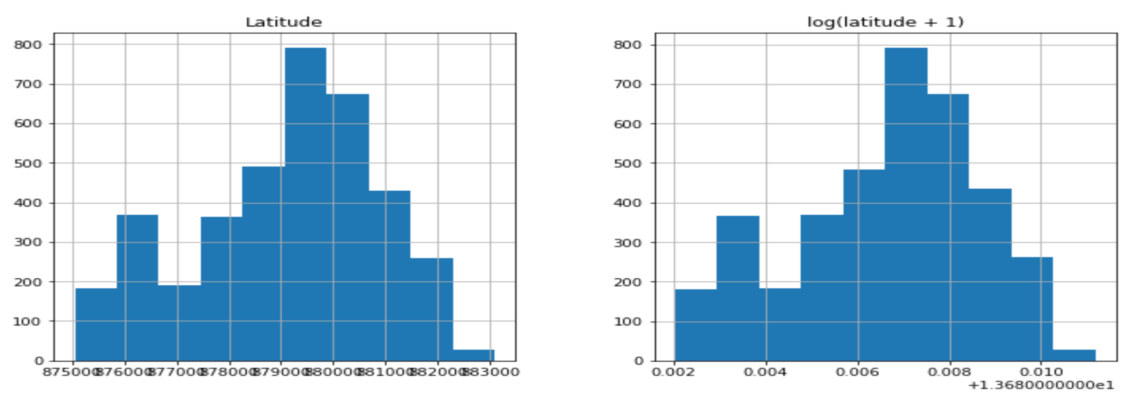



Figure 5.11 the features mutual information visual result

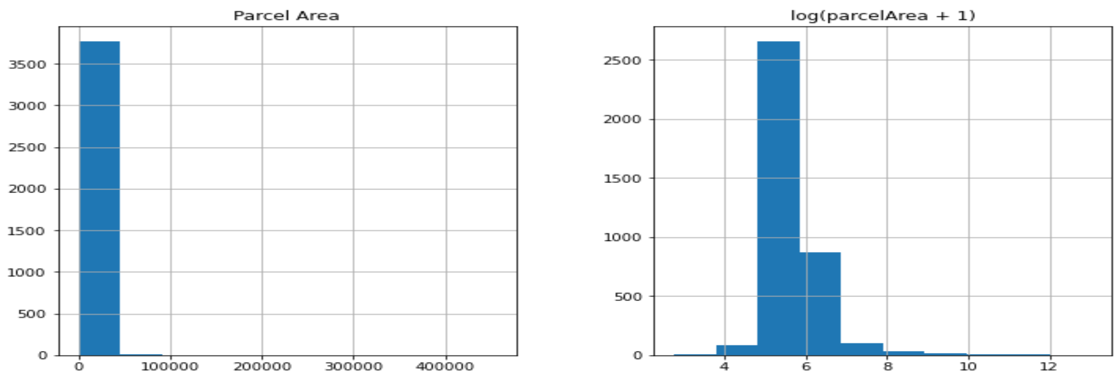
The data transformations are basic implementations to have transformed and normalized distribution of the features. We used logarithmic random exponential transformation to generate the normal distributed longitude, latitude and parcel area features because as we observe on figure 5.3 above these features have highest variance then other features. These transformations create attributes distribution of longitude features between 13.147 and 13.154, which was between 511884.6 and 516022.54. The distribution of latitude become between 13.68 and 13.69, which was between 874944.53 and 883086.62. The distribution of parcel area become between 3.044 and 6.217, which was originally between 15 and 454900, then reduced between 15 and 500 by mean ratio implementation as shown above on the figure 5.8. Fig 5.12 results show the transformed features before implementation of machine learning models to preprocessed dataset the features discussed.




***a. longitude skew-transformed feature***




***b. latitude skew-transformed features***




***c. parcel area skew-transformed feature***

Figure 5.12 the graphs of a, b and c are the visual results of power-transformed features

After feature transformation performed on the longitude, latitude, parcel area now we have normalized feature variables. Then we applied SelectBest class method from feature selection package to extract top ten features scores to the target variable. In addition, we concatenate two dataframe that are features and their score for better visualization. The figure 5.13 show the descending order result display of selectBest class scores.

| | Features | Score |
|---|---|---|
| 9 | fix | 575.539518 |
| 6 | mainRoadDistancekm | 243.257611 |
| 7 | riverDistancekm | 102.542608 |
| 5 | marketDistancekm | 26.910704 |
| 0 | kebele | 17.535134 |
| 8 | holdingType | 4.866536 |
| 3 | age | 1.999073 |
| 4 | parcelArea | 0.045663 |
| 2 | latitude | 0.000018 |
| 1 | longitude | 0.000001 |

Figure 5.13 the descending score of features to the target variable.

The dataset are being ready and we can select important features to determine our target. To determine the importance of the feature we used XGBClassifier() from xgboost package. Then we implement feature_importances_ package to the model. The fig 5.14 shows the visual result of important feature to determine the target.

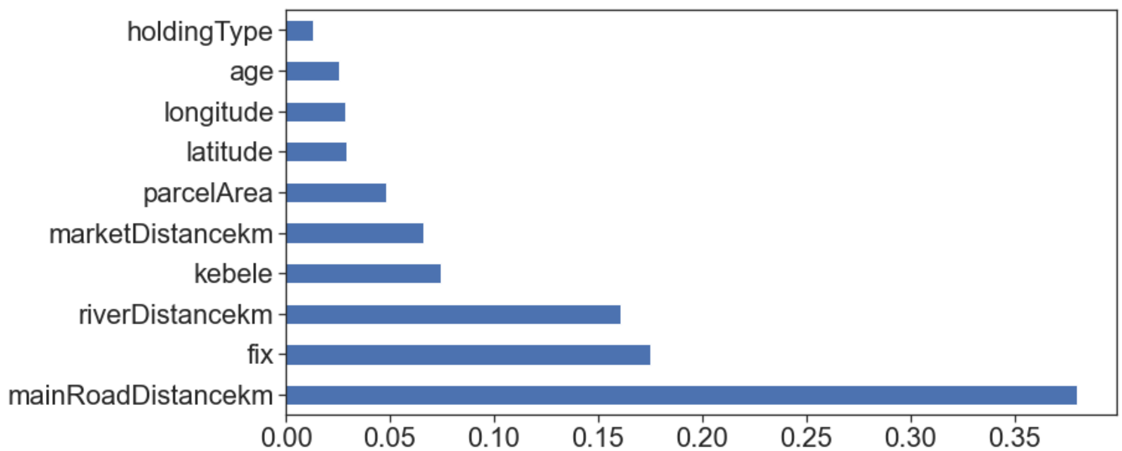
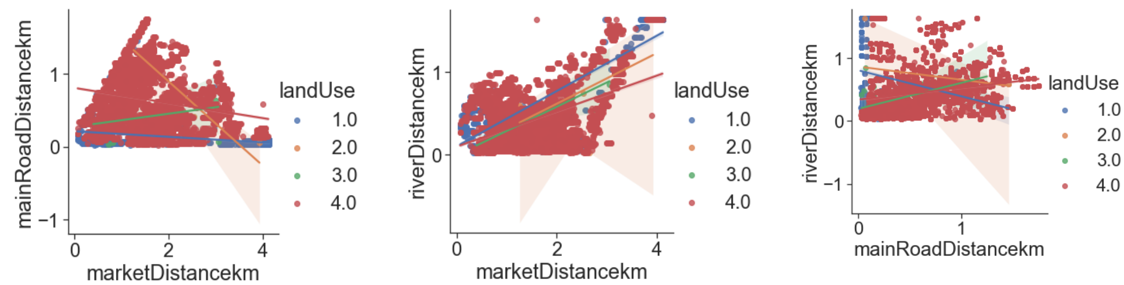



Figure **5.14** the visual result of important features.

From the result of feature importance score, the main attributes to determine target as (Snee, July 1983) state they easily identified checking the scores of features with random forest model. We plot an x-y line graph to show the frequency of data along a target line with the market, main road, and river distance along with each other because they are normal distributed features unlike fix and kebele features which are scaled between 0 and 1, and between 1 and 8 respectively. As figure 5.15 show that the market distance versus main road distance (fig 5.15 a) and market distance versus river distance (fig 5.15 b) has more saturated frequency than the main road versus river distance (fig 5.15 c). When we observe the linear relationship of the target line with these three graphs the market distance versus the river distance (fig 5.15 b) has a linear target line with less cross to each other than figure 5.15 a and figure 5.15 c display results. The graph (fig 5.15) shows that the main road, market and river distance are determinant of land uses and kebele also then another determinant of land uses in the Asella town. The following graph shows the x-y line graph of those three features of on the land uses.




a. features of market vs main road distance

b. features of market vs river distance

c. features of main road vs river distance

Figure **5.15** the graphs of a, b and c are of results the x-y graph of three features with the line target graph features.

## 5.2 Result of the study and discussion

For the experiment, we use imbalanced datasets, which result in five models for each dataset. Here the result of dataset balancing techniques and the result of classification models for dataset discussed. In the final section, we discuss the result obtained by each experiment in this study.

### 5.2.1 Result of dataset balancing techniques

To balance the selected class of the dataset the study used SMOTE resampling. The techniques applied to the dataset ready for modeling. The following Figures show the original and balanced dataset using SMOTE oversampling. As shown in Figures 5.16 is the commercial, 2 is industry, 3 is public and 4 is residential class of the dataset.

Figure 5.16 shows the original dataset with 7765 modules, from which 636 modules are commercial, 5 modules are industry, 27 modules are public and 7097 modules are residential. After balancing, the dataset shown in Figure 5.16 the all-class modules are 7097 class data with a total number of 28,388 data.

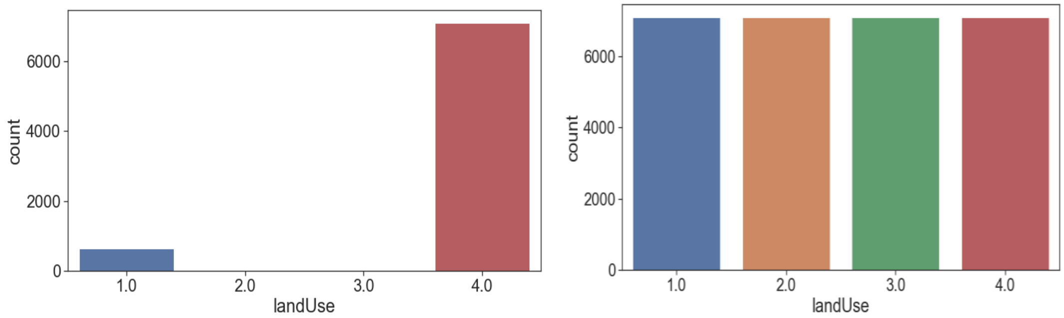



Figure **5.16** the visual count of class modules before and after balancing techniques on dataset

At the last StandardScaler used to scale the feature variable of prepared for model algorithms.

### 5.2.2 Model Evaluation Result

The performance of continues training using stratified cross validation of each models evaluated by accuracy, precision, recall, and F1-score are presented for KNN, RF, DT, SVM and NB respectively are shown in the table below.

The result in Table 5.1 shows the performance scores of the each model algorithm for dataset, the KNN model score the highest cross-validation accuracy with 99.24% and followed by RF which have 99.11% accuracy and the DT has also highest accuracy with 98.24% and the SVM scores 89.66% accuracy and lower accuracy recorded by NB model which is 86.37%. All model scored the best accuracy more than 86%. In addition, KNN, RF and DT tree performed with highest precision, recall and f1-score with 99%. The SVM and NB also score between 90% and 86% of precision, recall and f1-score.

**Table 5.1** Performance score for models

| Models | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| KNN | 99.24 | 99 | 99 | 99 |
| RF | 99.11 | 99 | 99 | 99 |
| DT | 98.24 | 98 | 98 | 98 |
| SVM | 89.66 | 90 | 90 | 89 |
| NB | 86.37 | 86 | 87 | 86 |

The bar chart in Figure 5.17 shows the visual representation of the all model evaluation of each model on the datasets according to the result of above in table. The blue represents the accuracy, the red represents precision, and the green represent recall and purple represent f1-scores of the models. The x-axis is represented the model and the y-axis is based on the accuracy result of the models. The accuracy, precision, recall and f1-score of KNN, RF and DT shows that the models have performed maximum performance of result for classify the target.

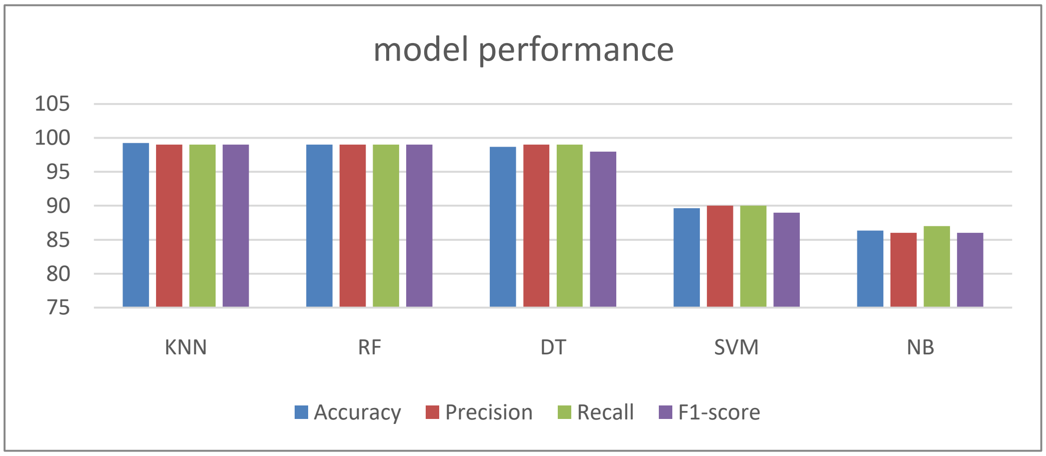



Figure **5.17** Model performance comparisons using each evaluation

### 5.2.3 Confusion matrix evaluation result

Confusion matrix is one way to evaluate the performance of classification models. The following figures are the normalize result of models.

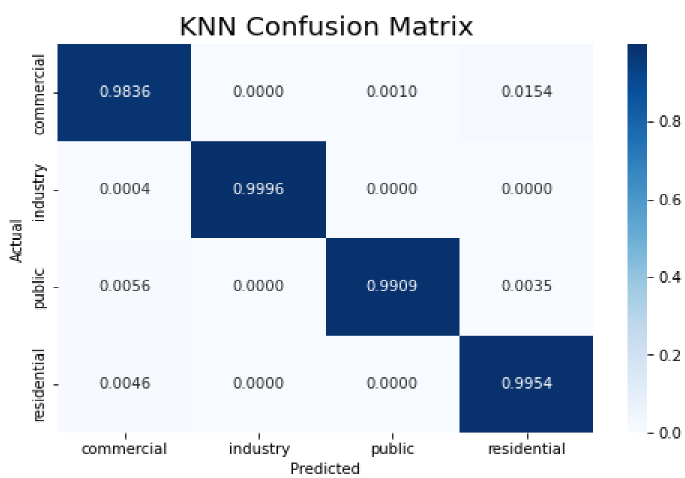



Figure **5.18** the confusion matrix of KNN model algorithm

Figure 5.18 shows the confusion matrix for KNN models. The KNN model classifies 98.36% of Commercial, 99.96% of Industry, 99.09% of Public, and 99.54% of Residential class correctly; but 1.54% of commercial misclassified as residential, 0.1% of commercial misclassified as public, 0.04% of industry misclassified as commercial, 0.56% of public misclassified as commercial, 0.35% of public misclassified residential and 0.46% of Residential misclassified as Commercial.

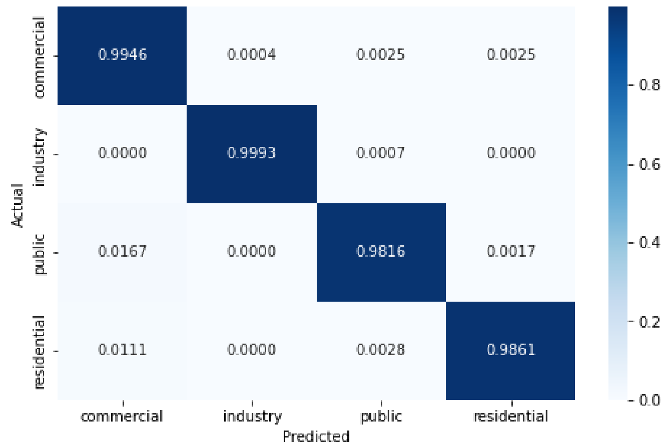



Figure **5.19** the Confusion matrix of Random Forest model algorithm

Figure 5.19 shows the confusion matrix for RF models. RF classifies 99.46% of commercial, 99.93% of Industry, 98.17% of Public and 98.61% of Residential class correctly; but 0.04%, 0.25% and 0.25% of Commercial misclassified as Industry, Public and Residential respectively. 0.07% of industry misclassified as public. 1.67%, and 0.17% of Public misclassified as Commercial and residential respectively. 1.11% and 0.28% of residential misclassified as commercial, industry and public respectively.

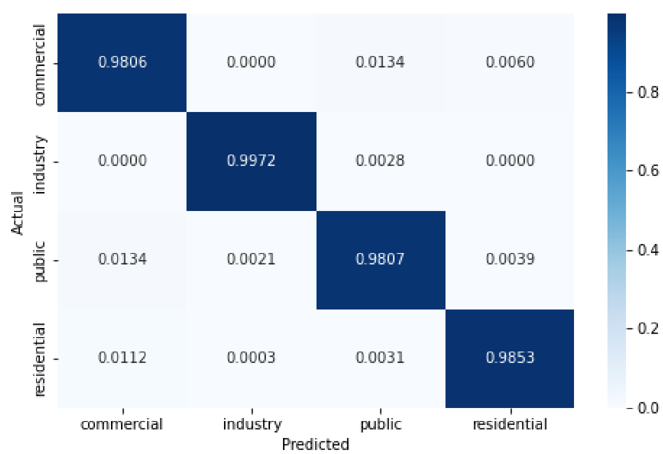



Figure **5.20** the Confusion matrix of Decision Tree model algorithm

Figure 5.20 shows the confusion matrix for DT models. DT classifies 98.06% of commercial, 99.72% of Industry, 98.07% of Public and 98.53% of Residential class correctly; but 1.34% and 0.04% of Commercial misclassified as Public and Residential respectively. 0.28% of industry misclassified as public. 1.34%, 0.21% and 0.39% of

Public misclassified as Commercial, Industry and residential respectively. 1.12%, 0.03% and 0.31% of residential misclassified as commercial, industry and public respectively.

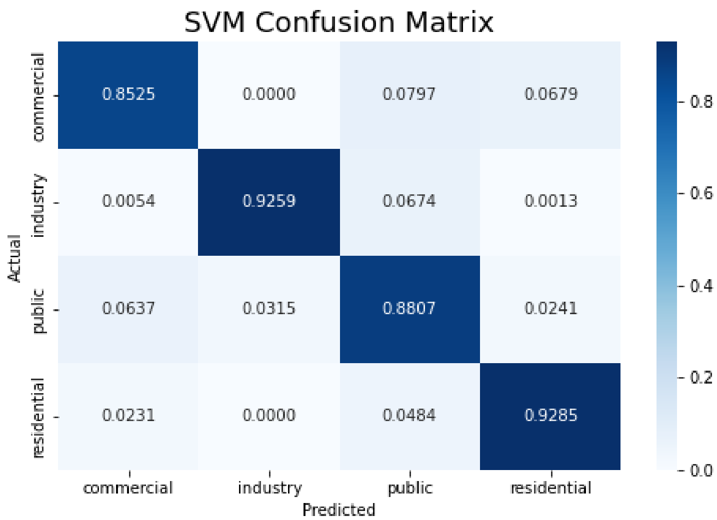



Figure **5.21** the confusion matrix of SVM model algorithm

Figure 5.21 shows the confusion matrix for SVM models. SVM on ELU classifies 85.25% of Commercial, 92.59% of Industry, 88.07% of Public, and 92.85% of Residential class correctly; but 7.97% and 6.79% of commercial misclassified as public and residential respectively, 0.54%, 6.74% and 0.13% of Industry misclassified as Commercial, public and residential respectively, 6.37%, 3.15% and 2.41% of public misclassified as commercial, industry and residential respectively. In addition, 2.31% and 4.84% of residential misclassified as commercial and public respectively.

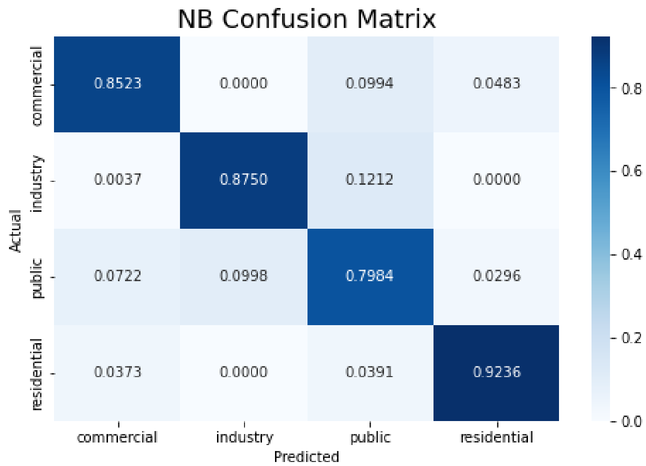



Figure **5.22** the Confusion matrix of NB model algorithm

Figure 5.22 shows the confusion matrix for NB models. NB classifies 85.23% of commercial, 87.50% of industry, 79.84% of public and 92.36% of residential class correctly; but 9.94% and 4.83% of commercial misclassified as Commercial and Residential. 0.37% and 12.12% of industry misclassified as commercial and public. 7.22%, 9.98% and 2.96% of public class misclassified as commercial, industry and residential classes respectively. In addition, 3.73% and 3.91% of residential class misclassified as commercial and public.

### 5.2.4 Discussion

We looked studied land-use prediction classification for land use and administration in this study. Aerial image detection changes used in certain research to classify land use and land cover levels. Our findings suggest that land use classification measures can be forecast land use classes.

The dataset we used for the study contains miss values, outliers and high difference of variance. The NaN values have not implemented whiles the imputation done on it. We replace the missed values of longitude, latitude, parcel area, built up area, market, main road and river distance features with mean values of the features and sex, age, holding method, title, holding type, PLU and land use features with mode values of the features. We also removed the detected outlier from longitude, market, main road and river distance features with IQR and Z-score. The mean ratio applied to parcel and built up area features. In addition, the features with high variance transformed using random logarithmic exponential transformation. Our dataset is highly unbalance features. Since the objective of the study is to predict classification of land use in the dataset target. We must deal with unbalance dataset to avoid the biasness of the model. To tackle the problem we used SMOTE over sampling to duplicate the number of minority classes. The result shows the method allows balancing without bias of classes

On the reclassification datasets, this study evaluates five machine-learning algorithms: KNN, RF, DT, SVM and BN. The dataset has 28,388 total data points with four land use classes. Continuous training accounts for predict the target and evaluate score of cross validation while equal class classification performed. For the dataset, the best accuracy reached by the KNN, RF and DT all performed almost above 99%. In addition, the NB model score the lowest accuracy with 86.37% compared to another model as shown in table 5.1.

The accuracy of KNN, RF and DT models score high efficiency with accuracy, precision, recall and f1-score. The SVM score decrease because of hyperplane, which used categories above, and under hyperplane of collected data. In case of our dataset, we have highly concentrated and little distributed dataset. These kinds of dataset are very suitable for KNN, RF and DT.  Because these model, use little similar parameters. KNN models collect each neighbor to determine the next target. Therefore, our dataset is very suitable for it. RF also uses bagging to estimate decorrelates tree, which use random picking rather than performing all training. Also concentrated dataset suitable for fandom forest.  DT also performs by hierarchy structure by starting node as a series branch, which suitable incase of highly concentrated dataset features.

# Conclusion and Recommendation

The suggested land use reclassification based on machine learning is discussed in this chapter, as well as the research findings. It also included a recommendation and study directions for the future.

## Conclusion

This study suggests using machine-learning techniques to solve developing a land-use prediction model. Machine learning approaches for reclassification land use were developed, implemented, and compared in this study. To carry out the study successfully, it was necessary to comprehend and establish land uses and prediction models, as mentioned in chapter two. Cleaning, balancing and cross validation using KNN, RF, DT, SVM and NB continuous model training and model testing and lastly comparing model using evaluation metrics are some approaches used to implement and land use classification prediction design.

To normalize each dataset, we cleaned and normalized it using cleaning, standardization and transform procedures. In the instance of dataset, each normalized dataset class grouped into land use categories Commercial, Industry, Public, and Residential. The model created utilizing KNN, RF, DT, SVM and NB based on these groups. The experiment conducted five models on the datasets.

In comparison to other classifiers, the KNN, RF and DT model had best greater than 99% cross validation score accuracy, precision, recall and f1-score. The study's findings were quite encouraging. When compared to the other models, the KNN, RF and DT models performed very well on dataset, hence utilizing the these model for land use prediction recommended.

## Recommendation and Future Work

Since the prediction, models have demonstrated some capacity to forecast the land use classes of land use modules. Consequently, the researcher suggests that land administration offices construct land use classification and distribution modules based on land use classification modules to improve land administration and management. Standard ecosystems and biodiversity also created because of these land use categories. To develop a model, the study advocated employing a clustering and classification machine learning technique. It is easier to see the differences in results when the models built using deep learning methods.

Future research can focus on constructing a model other characteristics such as priority and probability of a land-use classification for a more inclusive prediction model. Furthermore, no automated mechanism for online land use class has prediction addressed in the study's suggested land use prediction for land use modules on the online dataset. As a result, the research carried out using real-time land use classification.

## Bibliography


Abdi, A. M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GISCIENCE and REMOTE SENSING* , 1-20.

Abdul Fattah Mashat, M. M. (2012). A Decision Tree Classification Model for University. *International Journal of Advanced Computer Science and Applications* , 17-21.

Ayodele, T. O. (February 2010). *Types of Machine Learning Algorithms.* Portsmouth: University of Portsmouth.

Bao-wen, S. R. (2018). The research of regression model in machine learning field. *MATEC Web of Conferences 176, 01033.* YIN Chuan, China: https://doi.org/10.1051/matecconf/201817601033.

Barto, R. S. (2014, 2015). *Reinforcement Learning:An Introduction.* Cambridge, Massachusetts, London, England: The MIT Press.

Ben-David, S. S.-S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* New York: Cambridge University Press.

Berhanu Keno Terfa, N. C. (8 February 2020). Urbanization in Small Cities and Their Significant Implications on Landscape Structures: The Case in Ethiopia. *Sustainability 2020, 12, 1235; doi:10.3390/su12031235* .

Briassoulis, H. (2017). *Factors Influencing Land-Use and Land-Cover Change.* Mytilini, Lesvos, Greece : Encyclopedia of Life Support Systems.

Brownlee, J. (2016, March 16). *https://machinelearningmastery.com/.* Retrieved from Machine Learning Algorithms: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

C. Zhang, Y. W. (2019). Accelerated fading recognition for lithium-ion batteries with Nickel-Cobalt-Manganese cathode using quantile regression method. *Appl. Energy* , 113841.

Cahya, D. L., Martini, E., & Kasikoen, K. M. (2018). Urbanization and Land Use Changes in Peri-Urban Area using Spatial Analysis Methods (Case Study: Ciawi Urban Areas, Bogor Regency). *Earth and Environmental Science-2nd Geoplanning-International Conference on Geomatics and Planning* , Earth Environ. Sci. 123 012035.

Chekole, S. D. (2020). EVALUATION OF URBAN LAND ADMINISTRATION PROCESSES AND INSTITUTIONAL ARRANGEMENTS OF ETHIOPIA: BASED ON ADVOCACY COALITION THEORY. *African Journal on land policy and Geospatial Sciences* .

Christopher Bren d'Amoura, b. F.-H. (August 22, 2017 ). Future urban land expansion and implications for global croplands. *www.pnas.org/cgi/doi/10.1073/pnas.1606036114 PNAS* , 8939–8944.

COOK, A. (2022, January 13). *Scaling and Normalization*. Retrieved from https://www.kaggle.com/: https://www.kaggle.com/alexisbcook/scaling-and-normalization

CSA. (2021, Februry 03). *https://www.statsethiopia.gov.et*. Retrieved from https://www.statsethiopia.gov.et/: https://www.statsethiopia.gov.et/language/am/census/

D. Fecht, L. B. (2014). A gis-based urban simulation model for environmental health analysis. *Environmental Modeling & Software* , 1-11.

das, C. (2017, May 21). *What is machine learning and types of machine learning*. Retrieved from https://towardsdatascience.com: https://towardsdatascience.com/what-is-machine-learning-and-types-of-machine-learning-andrews-machine-learning-part-1-9cd9755bc647

Daud, S. D. (2021). Machine Learning Predictors for Sustainable Urban. *International Journal of Advanced Computer Science and Applications* , 772-780.

David Newton, R. P. (2018). RECENT TRENDS IN STOCHASTIC GRADIENT DESCENT FOR MACHINE LEARNING AND BIG DATA. *Proceedings of the 2018 Winter Simulation Conference* (pp. 366-380). Winter Simulation Conference: M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

De Soto, H. (2000). The Mystery of Capital: Why Capitalism Triumphs in the West and.

Di Gregorio, A. a. (1998). *Land cover classification system.* Rome: FAO.

Duhamel., C. (2011). *Land Use and Land Cover, Including Their Classification.* Luxembourg: Encyclopedia of Life Support Systems.

Eastman, D. (2020, August 12). *Buying Land the A to Z of Land Uses: Understanding Land-Use Specifics*. Retrieved January 10, 2022, from https://www.land.com/: https://www.land.com/buying/guide-to-land-use-definitions/

Enders, C. K. (2013). A Bayesian Approach for Estimating Mediation Effects with Missing Data. *Multivariate behavioral research* , 340-369.

Enemark, P. S. (2009). Land Administration Systems. *MAP WORLD FORUM.* HYEDRABAD, INDIA: Aalborg University, Denmark.

F. Karimi, S. S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems* , 61-75.

FIG. (February 1998). *Statement on the Cadastre.* Canberra, Australia: The International Federation of Surveyors (FIG).

Fischer, E. A. (1970). A Generalized k-Nearest Neighbor Rule. *Information and Control 16* , 128-152 .

Fisher, P. C. (2005). Land Use and Land Cover: Contradiction or Complement. *Re-Presenting GIS* , 85-98.

Flávio F. Camargo, E. E. (5 July 2019). A Comparative Assessment of Machine-Learning Techniques for Land Use and Land Cover Classification of the Brazilian Tropical Savanna Using ALOS-2/PALSAR-2 Polarimetric Images. *Remote Sens. 2019, 11, 1600; doi:10.3390/rs11131600* .

Garg, B. (June 2013). *DESIGN AND DEVELOPMENT OF NAÏVE BAYES CLASSIFIER.* Fargo, North Dakota: North Dakota State University.

Gilles Duranton, D. P. (2015). Handbook of Regional and Urban Economics. *Handbook of Regional and Urban Economics* .

Goldstein, J. R. (March 17, 2009). *Handling missing data.* Bristol city, UK: London School of Hygiene & Tropical Medicine / University of Bristol.

Guoyin Cai, H. R. (2019). Detailed Urban Land Use Land Cover Classification at the Metropolitan Scale Using a Three-Layer Classification Scheme. *www.mdpi.com/journal/sensors* , doi: 10.3390/s19143120.

Han J, P. J. (2011). *Data mining: concepts and techniques.* Amsterdam: Elsevier.

J. Yang, S. R. (2019). Outlier detection: How to threshold outlier? *Conf. Proceeding Ser.* (pp. 1-6). ACM Int.

James R. Anderson, E. E. (1976). *A Land Use and Land Cover Classification System for Use with Remote Sensor Data.* Washington: States Government Printing Office.

Jansen, L. (2006). Harmonization of land use class sets to facilitate compatibility and comparability of data across space and time. *Land Use Science Journal* , 127-156.

Jemal, M. Z. (January 2021). Software Defect Severity Level Prediction Using Machine Learning Techniques. *Partial Fulfillment of Requirement for the Degree* .

Johnson, M. K. (2013). *Applied Predictive Modeling.* New York: Springer, New York, NY.

KABACOFF, R. I. (2011). *R in Action.* Shelter Island, NY 11964: Manning Publications Co.

Kennedy, K. (2013). Credit Scoring Using Machine Learning. *Doctoral Science* .

Khan, A. B. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology* , 4-20.

*Land reform*. (n.d.). Retrieved Feburary 02, 2022, from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Land_reform

*Land Use*. (2022, Juanary 13). Retrieved from https://www.epa.gov/: https://www.epa.gov/report-environment/land-use

Lerrel Pinto, J. D. (2017). Robust Adversarial Reinforcement Learning. *International Conference on Machine.* Sydney, Australia: https://arxiv.org/abs/1703.02702v1.

Leta, M. K., Demissie, T. A., & Tränckner, J. (2021). Modeling and Prediction of Land Use Land Cover Change Dynamics Based on Land Change Modeler (LCM) in Nashe Watershed, Upper Blue Nile Basin, Ethiopia. *Sustainability 2021, 13, 3740* , https://doi.org/10.3390/su13073740.

Lewes, G. H. (January 2015). *Support Vector Machines for Classification.* American University of Beirut and Intel: https://www.researchgate.net/publication/300723807.

Livingston, F. (2005). Implementation of Breiman's Random Forest Machine Learning. *ECE591Q Machine Learning Journal Paper* , 1-13.

*Logistic Regression* . (2022, January 11). Retrieved from https://nhorton.people.amherst.edu/ips9/IPS_09_Ch14.pdf: https://nhorton.people.amherst.edu/ips9/IPS_09_Ch14.pdf

M. Ahmadlou, M. R. (2018). A comparative study of machine learning techniques to simulate land use changes. *Journal of the Indian Society of Remote Sensing* , 53- 62.

Mahesh Pal, P. M. (30 August 2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment* , 554-565.

Mather, M. P. (February 2002). *A Comparison of Decision Tree and Backpropagation Neural Network Classifiers.* Nottingham, U.K: ResearchGate.

Mitchell, T. M. (2006). *The Discipline of Machine Learning.* Pittsburgh: CMU-ML-06-108.

*ML | Evaluation Metrics*. (2021, August 08). Retrieved from https://www.geeksforgeeks.org/: https://www.geeksforgeeks.org/metrics-for-machine-learning-model/

*ML | Extra Tree Classifier for Feature Selection*. (2022, January 13). Retrieved from https://www.geeksforgeeks.org/: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/

Mulugeta, G. T. (2015). Addis Ababa City House Price Prediction by Using Regression Algorithm. *Performance Evaluation of Classifier Techniques to Discriminate Odors with an E-Nose* , 1-59.

N.E.Ulger, S. ,. (March 19-23, 2018). Land Use Problems and Land Management: A Land Inventory Study In Istanbul. *"World Bank Conference on Land And Poverty* (pp. 1-5). Washington DC: The World Bank.

Nesru H.Koroso, J. A. (December 2020). Urban land use efficiency in Ethiopia: An assessment of urban land use sustainability in Addis Ababa. *Land Use Policy* .

Nikola Kranjˇci´c, D. M. (18 March 2019). Support Vector Machine Accuracy Assessment for Extracting Green Urban Areas in Towns. *https://www.mdpi.com/journal/remotesensing* , 11/655 .

Nils Kok a, P. M. (2014). Land use regulations and the value of land and housing: An intra-metropolitan analysis. *Journal of Urban Economics* , 136-148.

office, A. T. (2007 -2012 E.C). *ASELLA SOCIO-ECONOMIC PROFILE.* Asella: Asella Town.

Oladipupo, T. (August, 2010). Types of Machine Learning Algorithms. *New Adv. Mach. Learn* , doi: 10.5772/9385.

Omeiza, D. (August 2019). *Efficient Machine Learning for Large-Scale Urban Land-Use Forecasting in Sub-Saharan Africa.* Carnegie Mellon University Africa: https://www.researchgate.net/publication/334866832.

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation, 8(6)* , 1-7.

Prasad, P. D. (May, 2012). Decision Tree Classification Model for Land Use and Land Cover Mapping a Case Study. *International Journal of Current Research* , 177-181.

Qingming Zhan, M. M. (2000). *Urban Land Use Classes With Fuzzy Membership and Classification Based on Integration of Remote Sensing and GIS.* Amsterdam : International Institute for Aerospace Survey and Earth Sciences (ITC).

Rajabifard, A., Daniel, S., & Ian, W. S. (n.d.). *Cadastral Templete*. Retrieved January 29, 2022, from http://cadastraltemplate.org/: http://cadastraltemplate.org/Cadastral Templete

RP, S. (2020, February 22). *Outlier Detection Techniques: Simplified*. Retrieved February 22, 2022, from https://www.kaggle.com: https://www.kaggle.com/rpsuraj/outlier-detection-techniques-simplified

S. García, S. R.-G. (2016). Big data preprocessing: methods and prospects. *Big Data Analaysis* , 1-22.

S. Tadese, T. S. (23 February 2021). Analysis of the Current and Future Prediction of Land Use/Land Cover Change Using Remote Sensing and the CA-Markov Model in Majang Forest Biosphere Reserves of Gambella, Southwestern Ethiopia. *□e Scientific World Journal* , 18 pages.

Sahalu, A. G. (February 2014). Analysis of Urban Land Use and Land Cover Changes: A Case Study in Bahir Dar, Ethiopia. *Partial fullfilment Degree of Master in Geospatial Technologies* .

Sarker IH, K. A. (2020). Cyber security data science: an overview from machine learning perspective. *J Big Data* , 1–29.

Serneels, S. E. (2006). Spatial sign preprocessing: a simple way to impart moderate robustness to multivariate estimators. *Journal of Chemical Information and Modeling* , 1402- 1409.

Shuo-sheng Wu, B. X. (July 2006 ). Urban Land-use Classification Using Variogram-based Analysis with an Aerial Photograph. *Article in Photogrammetric Engineering and Remote Sensing* , DOI: 10.14358/PERS.72.7.813.

Smilde, B. R. (2003). Centering and scaling in component analysis. *Journal of Chemometrics* , 16-33.

Snee, R. (July 1983). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Book Review. *Journal of Quality Technology* , 149-153.

Solomon Dargie Chekole, W. T. (2020). An Evaluation Framework for Urban Cadastral. *Land 2020, 9, 60; doi:10.3390/land9020060* , Land 2020, 9, 60.

SSCS, N. B. (2010). Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security* , 302-305.

Swapan Talukdar, P. S.-A. (2 April 2020). Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sens. 2020; https://doi.org/10.3390/rs12071135* , 12(7), 1135.

TONG, X. X. (2020). A Machine Learning-Based Method for Predicting Urban Land Use. *25th International Conference of the Association for Computer-Aided* (pp. 21-30). Hong Kong: Association for Computer-Aided Architectural Design Research in Asia (CAADRIA).

W. Mao, D. L. (31 August 2020). Comparison of Machine-Learning Methods for Urban Land-Use Mapping in Hangzhou City, China. *http://dx.doi.org/10.3390/rs12172817* .

Wei, Y. (August 1993). Urban Land Use Transformation and Determinants of Urban Land Use Size in China. *GeoJournal* , 435-440.

Weisberg, S. (2014). *Applied Linear Regression.* Minneapolis, MN: John Wiley & Sons, Inc., Hoboken, New Jersey.

Williams, H. A. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics; DOI: 10.1002/wics.101* , 433-459.

Wojtowytsch, S. (15 Sep 2021). Stochastic Gradient Descent with Noise of Machine Learning Type Part II: Continuous Time Analysis. *https://arxiv.org/abs/2106.02588v2* , arXiv:2106.02588v2 [cs.LG] 15 Sep 2021.

Wong, S. (2001). Internal Organization of Cities. *International Encyclopedia of the Social & Behavioral Sciences* .

*Z-score:Definition, Formula and Calculation*. (n.d.). Retrieved 02 21, 2022, from www.statisticshowto.com: www.statisticshowto.com/probability-and-statistics/z-score/

## Appendix A: Sample Source Code for Balancing

```
#resampling with SMOTE
from collections import Counter
from imblearn.over_sampling import SMOTE
smote = SMOTE()
# fit predictor and target variable
x_smote, y_smote = smote.fit_resample(x, y)
```

## Appendix B: Sample Source Code for modeling

```
#KNN
from sklearn.neighbors import KNeighborsClassifier
#define the parameters
knn_param= {'leaf_size':range(1,9),'n_neighbors': range(1,9), #default: 5
        'weights': ['uniform', 'distance'] #default = ‘uniform’
        }
#instanciate kNeighborsClassifier
knn_clf = KNeighborsClassifier()
#gridseearch
knn_model=GridSearchCV(knn_clf, knn_param)
# Create an instance of Pipeline Feature Scaling for input features
knn_scaled = make_pipeline(StandardScaler(), knn_model)
#evaluate cross value score of the model
score=cross_val_score(knn_scaled,x_smote,y_smote,scoring='accuracy',cv=skf)
knn_y_pred=cross_val_predict(knn_scaled,x_smote,y_smote,cv=skf)
print('KNN cross_val_scores = {:.2f}' .format(score.mean()*100, '%'))
# confusion matrix
conf_mat=confusion_matrix(knn_y_pred,y_smote)
# Normalise
cmn = conf_mat.astype('float') / conf_mat.sum(axis=1)[:, np.newaxis]
fig, ax = plt.subplots(figsize=((8, 5)))
sns.heatmap(cmn, annot=True, fmt='.4f', xticklabels=target_names,
yticklabels=target_names, cmap='Blues')
```

```
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.title('KNN Confusion Matrix', fontsize=18)
plt.show()
# classification report
knn_cr = classification_report(knn_y_pred,y_smote)
print('\n KNN Classification Report \n', knn_cr)
from sklearn.ensemble import RandomForestClassifier #RF
from sklearn.model_selection import GridSearchCV #GridSearch
rf_param = {'n_estimators': [10, 50, 100], #default=10
        'criterion': ['gini', 'entropy'], #default="gini"
        'max_depth': [3, 5, 10], #default=None
        'oob_score': [True], #default=False
        'random_state': [0]}
#instanciate random forest
rf_clf = RandomForestClassifier()
#gridSearch
rf_forest=GridSearchCV(rf_clf, rf_param)
# Create an instance of Pipeline
rf_scaled = make_pipeline(StandardScaler(), rf_forest)
#evaluate cross value score of the model
score=cross_val_score(rf_scaled, x_smote, y_smote,scoring='accuracy',cv=skf)
rf_y_pred=cross_val_predict(rf_scaled,  x_smote, y_smote, cv=skf)
print('cross_val_scores standardscaler= {:.2f}' .format(score.mean()*100, '%'))
target_names = ["commercial", "industry", "public", "residential"]
# confusion matrix
conf_mat=confusion_matrix(rf_y_pred,y_smote)
# Normalise confusion matrix
cmn = conf_mat.astype('float') / conf_mat.sum(axis=1)[:, np.newaxis]
fig, ax = plt.subplots(figsize=((8, 5)))
sns.heatmap(cmn, annot=True, fmt='.4f', xticklabels=target_names,
yticklabels=target_names, cmap='Blues')
```

```
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.title('RF Confusion Matrix', fontsize=18)
plt.show()
# classification report
rf_cr = classification_report(rf_y_pred, y_smote)
print('\n RF Cclassification Report \n', rf_cr)
# DT
from sklearn.tree import DecisionTreeClassifier
dt_param = {"max_depth":range(1,10),
        "min_samples_leaf": range(1, 9),
        "criterion": ['gini', 'entropy'] }
#initializing tree classifier
dt_clf= DecisionTreeClassifier()
#grid search
dt_model = GridSearchCV(dt_clf, dt_param)
# Create an instance of Pipeline
dt_scaled = make_pipeline(StandardScaler(), dt_model)
#evaluate cross value score of the model
score=cross_val_score(dt_scaled, x_smote,y_smote,scoring='accuracy',cv=skf)
dt_y_pred=cross_val_predict(dt_scaled,x_smote,y_smote,cv=skf)
print('DT cross_val_scores = {:.2f}' .format(score.mean()*100, '%'))
# confusion matrix
conf_mat=confusion_matrix(dt_y_pred,y_smote)
# Normalise
cmn = conf_mat.astype('float') / conf_mat.sum(axis=1)[:, np.newaxis]
fig, ax = plt.subplots(figsize=((8, 5)))
sns.heatmap(cmn, annot=True, fmt='.4f', xticklabels=target_names,
yticklabels=target_names, cmap='Blues')
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.title('DT Confusion Matrix', fontsize=18)
```

```
plt.show()
# classification report
dt_cr = classification_report(dt_y_pred,y_smote)
print('\n DT Classification Report \n', dt_cr)
#SVM
from sklearn.svm import LinearSVC
from sklearn.multiclass import OneVsRestClassifier
#inistanciate the training model
svc_model = LinearSVC(multi_class='ovr', max_iter=1000)
# Create an instance of Pipeline scaling the model.
svc_scaled = make_pipeline(StandardScaler(), svc_model)
#evaluate cross value score of the model
score=cross_val_score(svc_scaled,x_smote,y_smote,scoring='accuracy',cv=skf)
svc_y_pred=cross_val_predict(svc_scaled, x_smote, y_smote, cv=skf)
print('SVM cross_val_scores standardscaler= {:.2f}' .format(score.mean()*100, '%'))
# confusion matrix
conf_mat=confusion_matrix(svc_y_pred,y_smote)
# Normalise
cmn = conf_mat.astype('float') / conf_mat.sum(axis=1)[:, np.newaxis]
fig, ax = plt.subplots(figsize=((8, 5)))
sns.heatmap(cmn, annot=True, fmt='.4f', xticklabels=target_names,
yticklabels=target_names, cmap='Blues')
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.title('SVM Confusion Matrix', fontsize=18)
plt.show()
# classification report
svm_cr = classification_report(svc_y_pred,y_smote)
print('\n SVM Classification Report \n', svm_cr)
#NB
from sklearn.naive_bayes import GaussianNB
#instanciate Guassian Naive Bayes
```

```
nb_model = GaussianNB()
# Create an instance of Pipeline
nb_scaled = make_pipeline(StandardScaler(), nb_model)
#evaluate cross value score of the model
score=cross_val_score(nb_scaled,x_smote,y_smote,scoring='accuracy',cv=skf)
nb_y_pred=cross_val_predict(nb_scaled,x_smote,y_smote,cv=skf)
print('NB cross_val_scores standardscaler= {:.2f}' .format(score.mean()*100, '%'))
# confusion matrix
conf_mat=confusion_matrix(nb_y_pred,y_smote)
# Normalise
cmn = conf_mat.astype('float') / conf_mat.sum(axis=1)[:, np.newaxis]
fig, ax = plt.subplots(figsize=((8, 5)))
sns.heatmap(cmn, annot=True, fmt='.4f', xticklabels=target_names, yticklabels=target_names, cmap='Blues')
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.title('NB Confusion Matrix', fontsize=18)
plt.show()
# classification report
nb_cr = classification_report(nb_y_pred,y_smote)
print('\n NB Classification Report \n', nb_cr)
```