People’s Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Echahid Hamma Lakhdar University, El-Oued

Computer Science Department

Faculty of Exact Sciences

Title:

Robustness Analyzing of USSD Recharge Code by Using Artificial Intelligence Techniques

Presented by:

Slimane Gasmi

Ahmed Labadi

Supervisor:

Ammar Boucherit

Examination Committee:		
Chairman	Supervisor	Examiner
Ladjal Ibrahim	Ammar Boucherit	Bellila Khaoula

Academic Year: 2022/2023

تشكرات

أول مشكور هو الله عز وجل، ثم والداينا على كل مجهوداتهم منذ ولادتينا إلى هذه اللحظات، أنتم كل شيء أحبكم في الله أشد الحب.

يسرنا أن نوجه شكرنا لكل من نصحنا و أرشدنا أو ساهم معنا في إعداد هذا البحث بإيصالينا للمراجع والمصادر المطلوبة في أي مرحلة من مراحله، ونشكر على وجه الخصوص استاذنا الفاضل: عمار بوشريط على مساندتنا وإرشادنا بالنصح والتصحيح وعلى اختيار العنوان والموضوع، والشكر موصول لإدارة قسم الإعلام الآلي وكلية العلوم الدقيقة بجامعة الشهيد حمه لخضر.

إهداء

مرَّت قاطرة البحث بكثير من العوائق، ومع ذلك حاولت أن أتخطَّاها
بثبات بفضل من الله ومنِّه
إلى أبويَّ وأخوتي وأصدقائي، فلقد كانوا بمثابة العضد والسند في سبيل
استكمال البحث.
ولا ينبغي أن أنسى أساتذتي ممن كان لهم الدور الأكبر في مُساندتي
ومدِّي بالمعلومات القيِّمة ...أُهدي لكم بحث تخرُّجي داعيًا المولى - عزَّ
وجلَّ - أن يُطيل في أعماركم، ويرزقكم بالخيرات..

Contents

Abstract

USSD (Unstructured Supplementary Service Data) codes play a crucial role in various mobile services, including recharge transactions and interactive functionalities that involve sensitive data and financial transactions. Therefore, it is essential to identify and address any vulnerabilities or security weaknesses that could be exploited by malicious actors.

In this project, we will analyze the robustness of USSD recharge codes for a mobile phone operator in Algeria by using AI techniques to recognize patterns from historical data, with the aim to ensure their reliability, help detect vulnerabilities and weaknesses in the codes, enabling proactive measures to be taken.

The experimental results showed that the Artificial Neural Networks model performed better than the other models but in general the analyzed USSD recharge codes have a strong hidden pattern and, the mobile phone operator uses a good generating algorithm and possibly a continuous enhancement strategy in this context.

Key words: USSD, mobile services, AI techniques, Artificial Neural Networks.

الملخص

تلعب رموز USSD (بيانات الخدمة التكميلية غير المهيكلة) دورًا مهمًا في خدمات الهاتف المحمول المختلفة ، بما في ذلك معاملات إعادة الشحن والميزات التفاعلية التي تتضمن بيانات حساسة ومعاملات مالية. لذلك، من الضروري تحديد وإصلاح أي ثغرات أمنية أو نقاط ضعف يمكن استغلالها من قبل بعض الجهات لأغراض سيئة.

في مشروع التخرج الحالي، سنقوم بتحليل قوة رموز إعادة الشحن USSDلأحد شركات الهاتف المحمول الجزائرية باستخدام تقنيات الذكاء الاصطناعي للتعرف على الأنماط من البيانات المستعملة مؤخرا، بهدف ضمان موثوقيتها، والمساعدة في اكتشاف الثغرات ونقاط الضعف في هذه الرموز ، مما يتيح اتخاذ تدابير استباقية.

وقد أظهرت النتائج التجريبية أن نموذج والشبكة العصبية الاصطناعية كان أداؤه أفضل من النماذج الأخرى ولكن بشكل عام رموز إعادة شحن USSD التي تم تحليلها لها نمط مخفي قوي ، مما يوحي بأنّ مشغل الهاتف المحمول يستخدم خوارزميات توليد جيدة وربما استراتيجية تحسين مستمرة في هذا الإطار.

General Introduction

In recent years, the usage of mobile phones has become an integral part of our daily lives, and with it, the need for Unstructured Supplementary Service Data (USSD) recharge codes (prepaid cards) to top up phone credit. Recharge codes are numeric and/or alphanumeric strings that can be purchased and redeemed to add credit to a mobile phone account. The generation of these codes involves complex algorithms designed to ensure uniqueness, security, and reliability.

At same time, USSD recharge codes opened the doors to a large segment of fraudster customers, and which introduce a high damage to mobile phone operators themselves.

In this context, many people and academics believe it is impossible to correctly guess the 15 digits used to recharge a sim card. They claim that a complex algorithm is used to generate those numbers, making guessing nearly impossible.

However, we think that knowing the algorithm employed will speed up the process, but an attacker doesn’t need to know it to crack the numbers. Therefore, the reliability of these algorithms and the overall security of recharge codes remain a topic of interest and concern.

Research Problem

The vulnerability of USSD recharge codes poses a significant challenge in the domain of code security within the telecommunications industry; since such codes represent money and they are widely used to recharge mobile phone credit and perform various services, making them a prime target for potential exploitation.

In this context, the complexity of generating secure USSD codes lies in ensuring their uniqueness, resistance to unauthorized access, and protection against malicious activities such as code guessing or interception. The vulnerability of USSD recharge codes raises concerns regarding the potential compromise of financial losses.

Practically, understanding and addressing these vulnerabilities is crucial to safeguarding the integrity and reliability of recharge code systems, ensuring the trust and confidence of mobile phone users in the security of their transactions.

In other words, checking vulnerability of USSD codes by using machine learning will reassure users and help operators to enhance their mechanism's security for the generation of such codes.

This project serves as a culmination of our academic study, drawing upon the knowledge and skills acquired throughout our studies in computer science department. In this research, we seek to address the following research questions:

- What are the most known algorithms used in the generation of secure codes ?
- Is it possible to recognize the numeric pattern in USSD recharge code of the selected mobile operator by using machine learning techniques?

Research goals

In this study, we will delve into the domain of code analysis techniques, specifically focusing on the analysis of recharge code generation according to the most known algorithms used for secure code generation by using machine learning techniques.

Therefore, our primary objective is to evaluate the reliability and security of the algorithms used in generating these codes.

This project aims to shed light on the significance of code analysis techniques in evaluating the reliability of recharge code algorithms, thereby contributing to the improvement of code security and the overall functioning of telecommunication systems.

Thesis Structure

In the following, we will delve in the first chapter into the theoretical foundations related to our project. Then, chapter 2 recapitulates the basic concepts of machine learning and AI techniques. The third chapter is devoted to the research methodology, data gathering and selected machine learning models, and findings of our study. Furthermore, we will discuss the implications of our research and propose recommendations for the enhancement of recharge code generation algorithms in the general conclusion.

Chapter 1:
Basic Concepts on Mobile Networks

1. Introduction:

Mobile networks are quickly becoming the main type of network access for telecommunication services and they are one of the most developed and demandable transmittal usages [1]. In other words, mobile phone is one of the fastest growing and most demanding telecommunications applications. It represents a continuously increasing percentage of all new telephone subscriptions around the world.

In this context, there were around 6.6 billion smartphone subscribers worldwide in 2022. This number is forecast to exceed to 7.8 billion by 2028. The countries with the most smartphone mobile network subscriptions are China, India, and the United States [SW1].

On the other hand, the rivalry of making an increasing profit among the companies compels them to be inventive in terms of the services provided by the network and associated technology. Therefore, some services were started to be used like: Short Message Service (SMS), Unstructured Supplementary Service Data (USSD), Voice Call etc.

In this chapter we will introduce some basic concepts related to the subject of the study on mobile networks.

2. Basic Concepts and Information:

2.1. Mobile Networks:

A Mobile Network (MN) is a telecommunications network that provides services via radio signals for mobile phones. It is commonly described as a physical device that can be taken anywhere. In addition, a mobile network is designed to provide a cohesive and consistent framework for the development of a new generation of mobile data communication systems. This new generation of systems is necessary to meet the varied needs of current and future users, offering integrated support for the development, connection and operation of various mobile applications.

A mobile network is generally composed of four separate parts that work together to function as a whole: the mobile device itself, the base station subsystem (BSS), the network switching subsystem (NSS) and the operation and support subsystem (OSS). The following figure Fig.1.1 illustrates this architecture and more details can be found in [2].

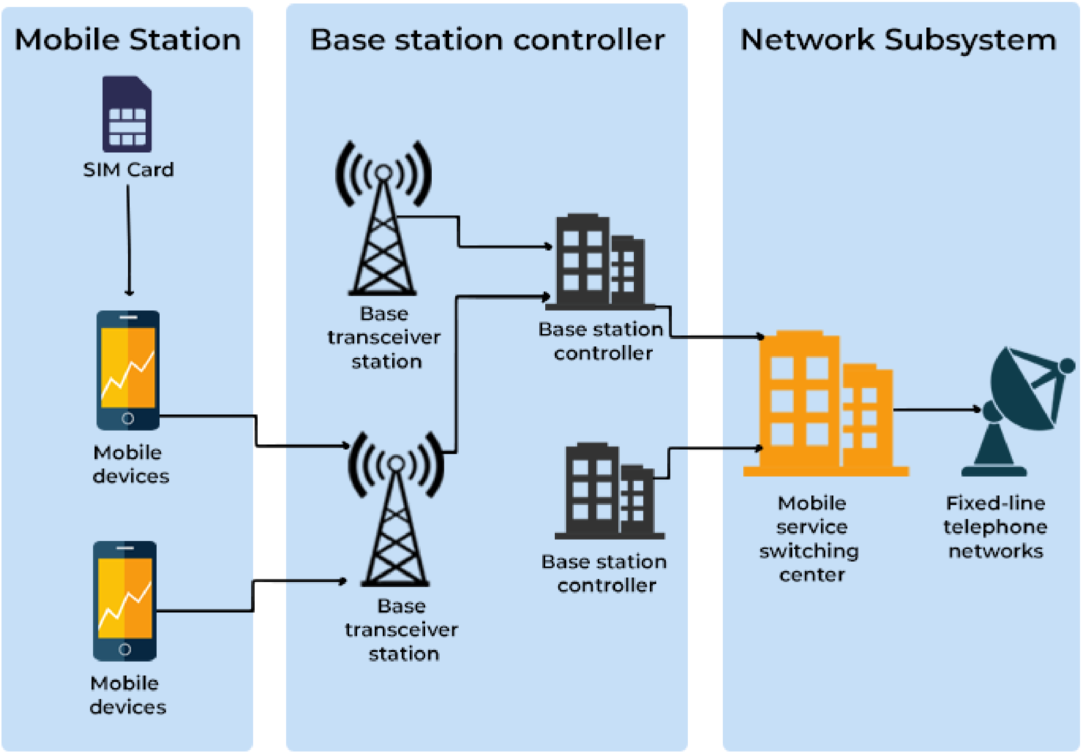

Fig.1.1. Mobile network composition [SW4]

Although Mobile networks were designed as a secure wireless system, they can still be subject to attacks. Mobile networks use authentication measures, such as challenge-response authentication, which prompts a user to provide a valid answer to a question, and a pre-shared key which comes in the form of a password or passphrase.

2.2. Algerian Mobile Networks:

According to a new report made by the Post and Electronic Communications Control Authority, there are about 47.67 million active mobile phone subscribers were registered in Algeria during the first quarter of 2022, compared to 46.04 million subscribers in the first quarter of 2021, an increase of 3.35 %. The same source stated that out of 47.67 million active subscribers, 42.68 million are subscribers to the third and fourth generation networks, or 89.54 percent, compared to 4.98 million subscribers to the "GSM" network, i.e. 10.46 percent.

In the same context, the mobile operator Mobilis maintained the first rank in terms of the number of subscribers in the GSM networks for the third and fourth generation during the first quarter of 2022, with 20.3 million subscribers, followed by Djeezy (14.6 million) and Ooredoo (12. 7 million).

The Mobilis customer recorded a development in the fold of its GSM, 3rd and 4th generation subscribers, with 20.3 million subscribers during the current quarter, compared to 19.2 million during the same period last year [SW2].

Another report [SW3] indicates that mobile phone connections in Algeria were equivalent to 107.2% of the total population in January 2023, and this indicates that a number of people in Algeria use more than one mobile phone to connect to the Internet. As the statistics show that mobile Internet use in Algeria is increased by 1.8 million (+3.8 percent) between 2022 and 2023. The following figure Fig.1.2. illustrates the increasing percentage of mobile phone subscriptions in comparison with fixed phone subscriptions.

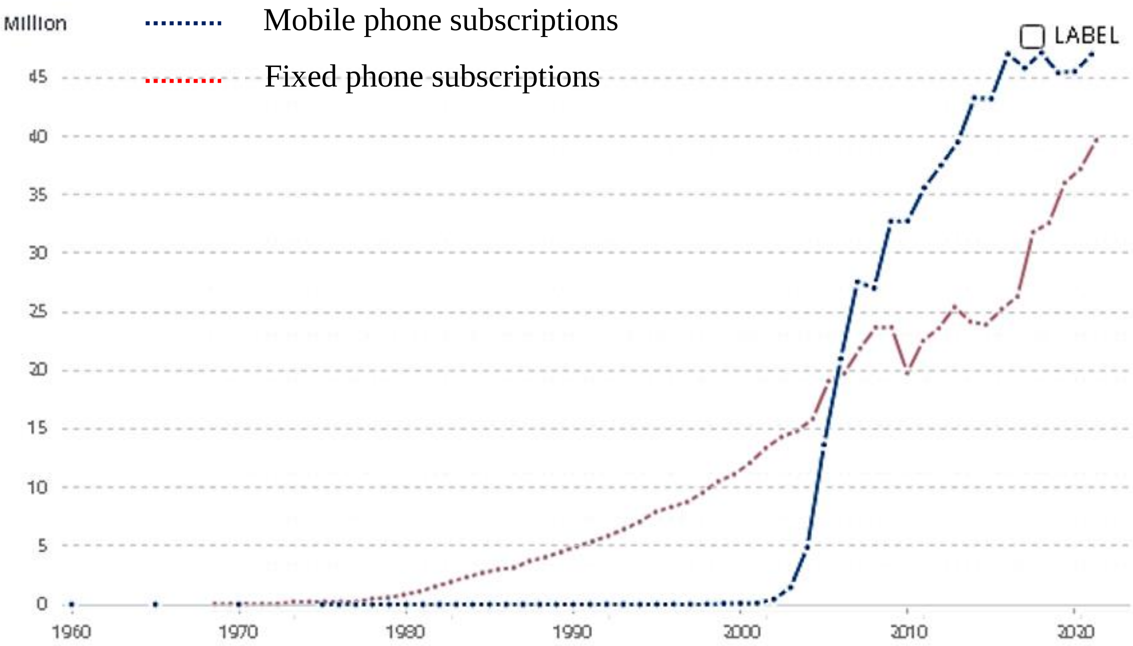

Fig.1.2. Mobile and fixed phone subscription growth in Algeria [SW5]

2.3. Mobile Network services:

In addition to the basic services, this subsection is devoted to provide a brief description of the most popular services such as [3]:

1. **Short Message Service (SMS)** : is a basic service allowing mobile stations and other network-connected devices to exchange short text messages. The first short text message had been transferred in 1992 over signaling channels of a European GSM network. Since this successful experience, SMS usage has been the subject of exponential growth.

2. **Multimedia Messaging Service (MMS)** : allows users of mobile phones connected to the cellular mobile telephone network to send images, audio files, and text messages. Almost all GSM service providers offer this service, offering customers the opportunity to create new business models based on this technology.

3. **Unstructured Supplementary Service Data (USSD)** : is a text based messaging protocol used in GSM to establish communication between mobile phones and the operators' special application servers over signaling channels. In such a service, a transfer of up to 182 alphanumeric characters is supported.

In contrast to SMS messages, USSD messages create a real-time connection during a USSD session. The connection that allows a two-way exchange of a sequence of data, remains active until a termination message is sent or a timeout is reached. In addition, USSD is almost seven times faster than SMS. Moreover, USSD can be used by the user to send command to an application in the text format [4]. The following figure Fig.1.3. illustrates the work principle of USSD.

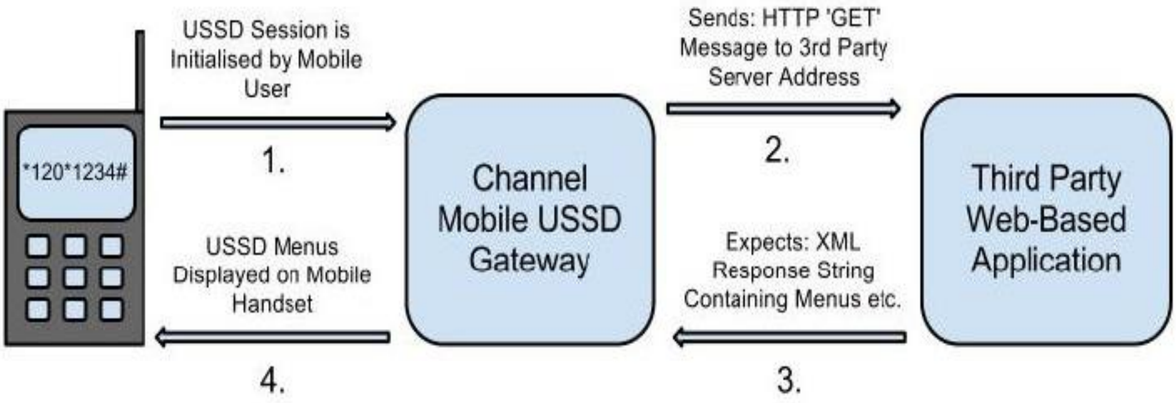
USSD Session is Initialised by Mobile User
1.
Sends: HTTP 'GET' Message to 3rd Party Server Address
2.
*120*1234#
Channel Mobile USSD Gateway
Third Party Web-Based Application
USSD Menus Displayed on Mobile Handset
4.
Expects: XML Response String Containing Menus etc.
3.

Fig.1.3. USSD work principle

3. USSD recharge codes:

Today, USSD is used by banks, financial service providers, the government, and many other stakeholders to communicate and receive information from mobile phone users. Although USSD is used so often in daily life, it is not known that the service which is used while entering the prepaid card's code by entering *111* card's code # for instance is USSD. In this subsection we give a simple description of a voucher (prepaid card) based recharges.

3.1 Voucher recharge

In such recharge system, a phone user dials a USSD short code which includes voucher activation code (eg: *111*373538996243901#). Then if the voucher code is correct, the account is credited with recharge amount. Otherwise, an error message is generated. A simple description of USSD based voucher recharge is given as follow:

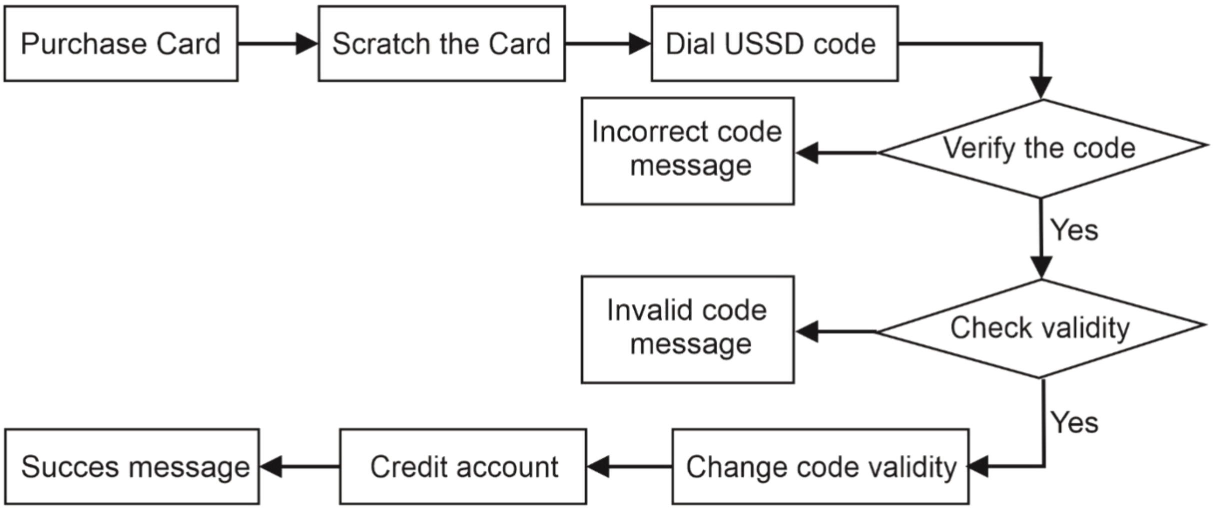

Fig.1.4. USSD based voucher recharge steps

The main steps given in figure Fig.1.4. are described as follow [5]:

1. Subscriber purchases, scratches and dials the USSD short code with voucher code. The request goes to the BTS (Base Transceiver System) that forwards the request to the BSC (Base Station Subsystem).

2. BSC forwards the request to MSC (Mobile Service Switching system) which does some verifications (format, etc). Then, forwards the request to HLR (Home Location Register) which understands the USSD code that it's a recharge request.

3. HLR forwards the recharge request to Recharge handling system of IN (Intelligence Network). The IN holds the subscriber database including different balances, tariffs, recharge rules, charging rules etc.

4. Recharge handling component sends the voucher code which was received from subscriber to Voucher Server to check it's value and also to change it's status to "USED". If the voucher is already in USED state, then recharge will fail and send an error message.

5. Recharge handling system will update the balance in the SDP (Subscriber Data Point). SDP is the component of the IN which holds all the subscriber's account details including balance, life-cycle etc.

6. SDP responds with success or failure for the update request. The recharge response is sent back to subscriber.

3.2. Voucher code generation:

The Voucher Server is the system where voucher pins are generated and stored. On the Voucher Server, the administrator will provide the Transaction Amount, Expiry, Serial Number Range, Agent Details etc. as an input and an algorithm present on system will generate the voucher codes. Generally, these codes are encrypted.

When the vouchers are created, they are in “Generated” or “Created” state. These pins cannot be used for recharges unless they are in “AVAILABLE” state. So, a state change is performed on these vouchers are per requirement (usually before printing them) to make them in “AVAILABLE” state. Once that is done, the vouchers are ready to use.

Finally, when the user performs the recharge, the voucher state is change to “USED” and it cannot further be used to refill the account. Such description is just to give a high level idea on how the voucher pins are generated and how recharge works [6].

4. Codes verification and validation algorithms

There are some known algorithms that are designed to ensure the validity of credit and/or prepaid cards and that they have not been used before (case of prepaid cards). This involves several steps, such as verifying the card number (digits), expiration date, and security code (CVV/CVC), as well as checking the card's status with the issuing bank or payment network. In this section we present some of them without entering in details.

4.1. Luhn Algorithm:

The Luhn algorithm also known as (MOD 10) is a simple checksum formula used as first line of defense in many e-commerce and governmental sites to validate a variety of identification numbers, such as credit cards, IMEI numbers, National Provider Identifier numbers in United states, Canadian Social Insurance Numbers etc. The last digit of the Luhn based code is called a check digit.

The Luhn algorithm formula was created by German Computer Scientist Hans Peter Luhn while he is working at IBM as a researcher and created this validator using modern mathematical algorithms that enabled computers to ascertain the correct input of identification numbers quickly.

This checksum formula is widely used today, especially to protect companies and consumers against accidental errors and typing errors [7].

Principle:

Based on this algorithm, the process of verification includes the following steps:

1) Multiply every even-position digit (starting from right) by two and let others as they are.

(a) If the product is equal or greater than ten, add the digits to produce one digit.

2) calculate the total of all the resultants (even-position) that we have in step 1.

3) calculate the total of all the numbers in the odd-positions, except the rightmost digit.

4) Add these last two values together.

5) The rightmost digit should be what it takes to get the sum to be a multiple of 10.

Example:

Let us consider the following credit card number sequence ‘09748649825741X.’

Suppose that we would like to determine the correct value of ‘x’ using Luhn’s algorithm.

Squence	**0**	**9**	**7**	**4**	**8**	**6**	**4**	**9**	**8**	**2**	**5**	**7**	**4**	**1**	**X**
Step 1:	0	18	7	8	8	12	4	18	8	4	5	14	4	2	
		9		8		**3**		**9**		4		**5**		2	
Step 2:	Total of evens = 40														
Step 3:	Total of odds = 36														
Step 4:	Total = 40+36=76														

Step 5 : Check digit is = (10 – (76 mod 10)) mod 10 = 4

Thus, the value of X and the complete credit card number is ‘09748649825741**4**’

4.2. Universal Product Code :

A Universal Product Code (UPC) is a code on a product's packaging that helps in product identification. It has a black barcode and a 12-digit number under it. UPC barcodes are scannable symbols that are made to make it easier to identify product characteristics, such as brand name, size, item, size, or color. The last digit of the UPC code is also called a check digit.

UPCs have a number of advantages to businesses and consumers. Because they make it possible for barcode scanners to immediately identify a product and its associated price, UPCs improve speed, efficiency, and productivity by eliminating the need to manually enter product information [8].

Principle:

Based on UPC, the check digit can be calculated following these steps:

1) Multiply all odd-position digit (starting from left) by three and let others as they are.

2) calculate the total of all the resultants (odd-position) that we have in step 1.

3) calculate the total of all the numbers in the even-positions, except the rightmost digit.

4) Add these last two values together.

5) The check digit is the number that, when added to the number in step 4, is a multiple of 10.

Example:

Let us consider the following UPC barcode ‘63233500079X.’

Suppose that we would like to determine the correct value of ‘x’ using UPC code.

Squence	**6**	**3**	**2**	**3**	**3**	**5**	**0**	**0**	**0**	**7**	**9**	**X**
Step 1:	**18**	3	**6**	3	**9**	5	**0**	0	**0**	7	**27**	
Step 2:	Total of odds = 60											
Step 3:	Total of evens = 18											
Step 4:	Total = 60+18= 78											

Step 5 : Check digit is = (10 – (78 mod 10)) mod 10 = 2

Thus, the value of X and the complete the given UPC barcode is ‘63233500079**2**’

4.3. Codabar:

The codabar was developed in 1972 by Pitney Bowes with the intention of using it initially in the retail merchandise industry. However, its professional applications soon extended to medical and educational institutions and shipping companies.

The codabar barcodes were designed to overcome difficulties faced when using Monarch Code and to be accurately read even when printed using dot-matrix printers and typewriter-like impact printers. As a result, this barcode symbology gained popularity in multi-part forms like FedEx airbills and blood bank forms. Today, Codabar barcodes are commonly used in the library industry to facilitate the organizing and tracking of books.

A Codabar barcode can encode numeric digits ranging from 0-9 and five special characters, including Plus (+), Minus (-), Forward Slash (/), Colon (:), Dollar symbol ($), and Dot (.) [SW7].

Principle:

The check digit in Codabar barcode can be calculated as follows:

As the character of Codabar are: 0123456789-$:/.+ therefore, they have orders from 0 to 15.

1) attribute to each digit the corresponding order.

2) calculate the total of all character orders in the Codabar code obtained in step 1.

3) The check digit is the number that, when added to the number in step 4, is a multiple of 16.

Remark:

The Codabar code can be used only with decimal digits, therefore, the check digit is calculated in order to have a multiple of 10.

Example:

Let us consider the following Codabar barcode ‘73235107900X.’

Suppose that we would like to determine the correct value of ‘x’ using Codabar barcode.

Squence	**7**	**3**	**2**	**3**	**5**	**1**	**0**	**7**	**9**	**0**	**3**	**X**
Step 1:	7	3	2	3	5	1	0	7	9	0	3	
Step 2:	Total of orders = 40											

Step 3 : Check digit is = (16 – (40 mod 16)) mod 16 = 8

Thus, the value of X and the complete the given Codabar barcode is ‘73235107903**8**’

5. Conclusion:

In this chapter, we talked about the basic concepts we need to know about USSD recharge codes as well as how such codes are generated. In addition, we have presented three well-known algorithms used for the verification and validation of codes. In the next chapter, we will present our approach for the analysis of USSD recharge codes.

Chapter 2:

Basic Concepts on IA techniques

1. Introduction:

Utilizing sample data or prior knowledge, machine learning involves programming computers to optimize a performance criterion. We have a model that is defined up to a certain point, and learning is the process of running a computer program to optimize the model's parameters using training data or previous knowledge. The model could be descriptive to learn from the data or predictive to make future predictions.

The phrase "Machine Learning" was first used in 1959 by Arthur Samuel, an early American pioneer in the fields of artificial intelligence and computer gaming. Samuel was working for IBM at the time. "The field of study that gives computers the ability to learn without being explicitly programmed," he said of machine learning. There isn't a single definition of machine learning, though. various authors.

2. Machine Learning :

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E.

Examples:

I- Handwriting recognition learning problem

- Task T: Recognizing and classifying handwritten words within images
- Performance P: Percent of words correctly classified
- Training experience E: A dataset of handwritten words with given classifications

II- A robot driving learning problem

- Task T: Driving on highways using vision sensors
- Performance measure P: Average distance traveled before an error
- training experience: A sequence of images and steering commands recorded while observing a human driver

III- A chess learning problem

- Task T: Playing chess
- Performance measure P: Percent of games won against opponents
- Training experience E: Playing practice games against itself

Definition:

A computer program which learns from experience is called a machine learning program or simply a learning program. Such a program is sometimes also referred to as a learner.

2.1. How machines could learn?

2.1.1 Basic components of learning process:

The learning process, whether by a human or a machine, can be divided into four components, namely, data storage, abstraction, generalization and evaluation. figure Fig.2.1 illustrates the various components and the steps involved in the learning process.

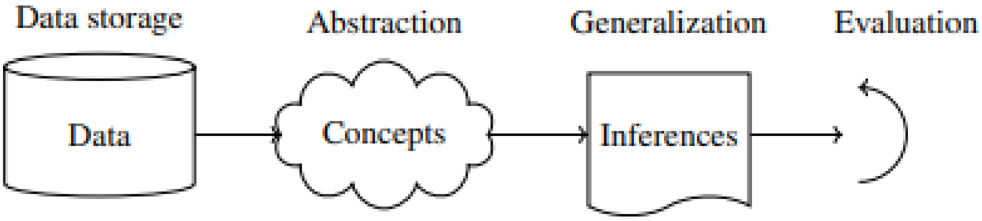

Fig.2.1. components of learning process

1. Data Storage:

The storage and retrieval of large amounts of data are vital components of the learning process. Both humans and computers rely on data storage to support advanced reasoning.

- Humans store data in the brain and retrieve it through electrochemical signals.

- Computers utilize devices such as hard disk drives, flash memory, and random access memory (RAM) to store data. They employ various technologies, including cables, to retrieve the stored data.

2. Abstraction:

Abstraction constitutes the second element of the learning process. It involves extracting knowledge from stored data by creating general concepts that represent the data as a whole. This process incorporates applying existing models and developing new ones.

- Training is the process of fitting a model to a dataset. Once the model is trained, the data is transformed into an abstract representation that summarizes the original information.

3. Generalization:

Generalization is the third component of the learning process. It refers to the process of utilizing knowledge derived from stored data for future actions. These actions pertain to tasks that are similar, though not identical, to those encountered previously. The objective of generalization is to identify the properties of the data that are most relevant to future tasks.

4. Evaluation:

Evaluation serves as the final component of the learning process. It involves providing feedback to measure the usefulness of acquired knowledge. This feedback is then used to improve the entire learning process.

3. Main types of learning:

In general, machine learning algorithms can be classified into three types.

3.1. Supervised learning:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.

In supervised learning, each example in the training set is a pair consisting of an input object (typically, a vector) and an output value. A supervised learning algorithm analyzes the training data and produces a function, which can be used for mapping new examples. In the optimal case, the function will correctly determine the class labels for unseen instances. Both classification and

regression problems are supervised learning problems.

A wide range of supervised learning algorithms are available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems.

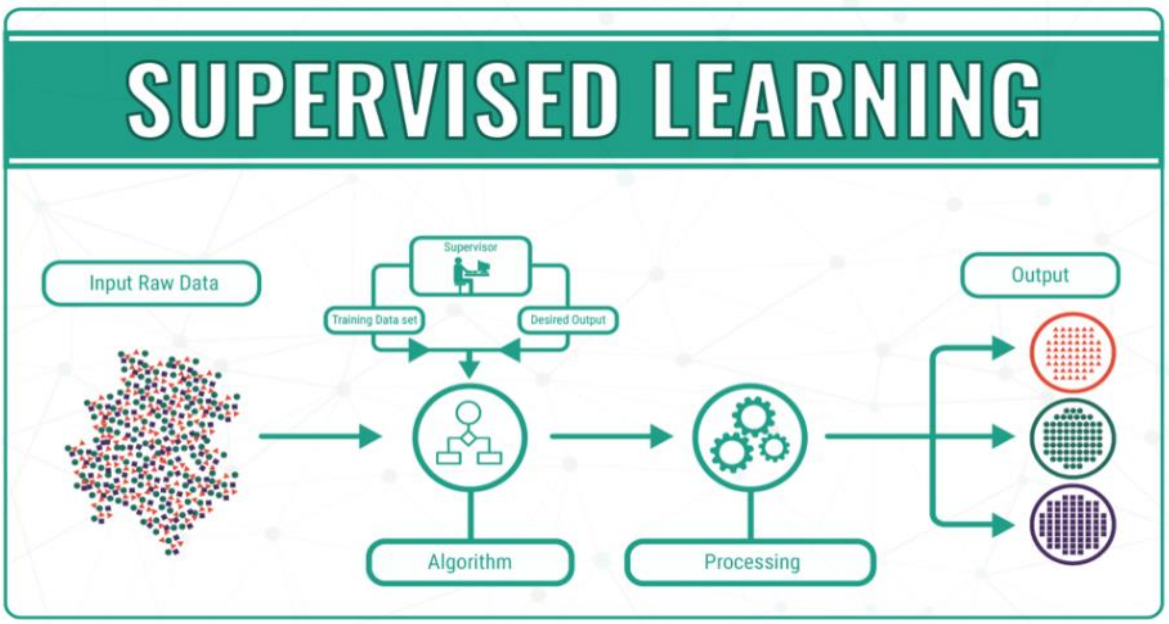

Fig.2.2. Supervised learning

Remarks:

The term "supervised learning" is used because the algorithm's learning process can be likened to being supervised by a teacher. In this type of learning, the algorithm is provided with a training dataset that includes the correct answers or outputs. It makes predictions on the training data and receives corrections from the teacher. This iterative process continues until the algorithm reaches a satisfactory level of performance, at which point the learning process is halted.

Examples:

1. Linear Regression:

This algorithm is used for predicting a continuous output variable based on one or more input features. It fits a linear equation to the training data to establish a relationship between the inputs and the target variable.

2. Logistic Regression:

It is used for binary classification problems, where the goal is to predict one of two classes. Logistic regression estimates the probability of an input belonging to a particular class using a logistic function.

3. Decision Trees:

Decision trees build a hierarchical structure of if-else conditions based on the input features to make predictions. They partition the input space into regions, assigning class labels to each region.

4. Random Forest:

Random Forest is an ensemble method that combines multiple decision trees to make predictions. It generates a set of decision trees and uses averaging or voting to determine the final prediction.

These are just a few examples of supervised machine learning algorithms. Each algorithm has its strengths and limitations, and the choice depends on the specific problem and dataset at hand.

3.2. Unsupervised Learning

Unsupervised learning refers to a category of machine learning algorithms that analyze datasets without labeled responses. In contrast to supervised learning, unsupervised learning does not involve classifying or categorizing the observations. There are no provided output values, and thus, there is no estimation of functions based on labeled data. Because the examples given to the learning algorithm are unlabeled, it is not possible to evaluate the accuracy of the resulting structure.

Cluster analysis is the most common technique used in unsupervised learning. It is employed in exploratory data analysis to uncover hidden patterns or groups within the data.

Examples:

1. K-Means Clustering:

K-means clustering is a popular algorithm for grouping similar data points into a predefined number of clusters. It iteratively assigns each data point to the nearest cluster centroid and updates the centroids until convergence.

2. Hierarchical Clustering:

Hierarchical clustering builds a tree-like structure (dendrogram) to represent the relationships between data points. It can be agglomerative (bottom-up) or divisive (top-down) and does not require a predetermined number of clusters.

3. Principal Component Analysis (PCA):

PCA is a dimensionality reduction technique that identifies the most important features in a dataset. It transforms the data into a new set of orthogonal variables called principal components, which capture the maximum amount of variance.

4. Apriori Algorithm:

Apriori is a classic algorithm for discovering frequent itemsets in transactional datasets. It identifies associations or patterns among items by determining which items tend to appear together in transactions.

These are just a few examples of unsupervised machine learning algorithms. Unsupervised learning is often used for exploratory data analysis, pattern discovery, and data compression, among other tasks. The choice of algorithm depends on the specific problem and the nature of the data.

3.3. Classification and Regression:

3.3.1. Classification:

A classification tree is an algorithm specifically designed to handle fixed or categorical target variables. Its purpose is to determine the most likely class or category to which a given target variable belongs. For instance, in the context of predicting subscription to a digital platform or high school graduation, these can be considered as simple binary classification problems where the dependent variable can take only one of two exclusive values.

However, there are scenarios where the categorical dependent variable can have multiple possible values. For example, predicting the type of smartphone a consumer is likely to purchase may involve multiple categories. In such cases, a classic classification tree is used to create a visual representation of the decision-making process.

A classification tree visually depicts a series of decision nodes and branches, representing different features or variables that lead to specific class predictions. Each node represents a test on a particular feature, and the branches emanating from the node correspond to the possible outcomes of that test. The terminal nodes, also known as leaves, provide the final class predictions based on the path followed through the decision tree.

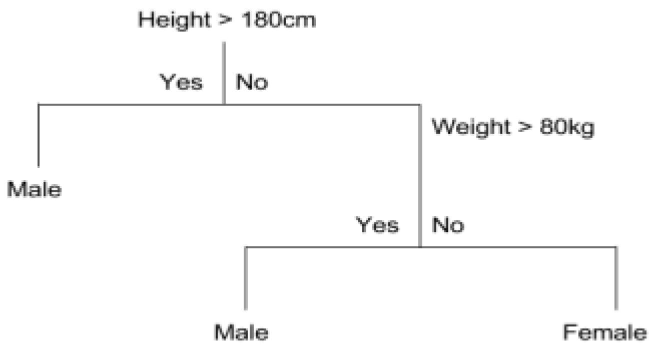

Fig.2.3. example of classification

3.3.2. Regression:

A regression is an algorithm designed to predict the value of a target variable, where the target variable is continuous. In regression-type problems, the goal is to estimate or forecast numerical values, such as predicting the selling prices of residential houses.

When using a regression, the algorithm takes into account both continuous factors (e.g., square footage) and categorical factors (e.g., style of home, area of the property) that may influence the target variable. These factors are used as inputs or features in the tree, and the algorithm recursively splits the data based on different thresholds or conditions to create a hierarchical structure.

At each node of the regression, a feature is selected to create a split that optimally separates the data based on the target variable. The splitting process continues until certain stopping criteria are met, such as a maximum depth or a minimum number of data points in each leaf node.

Ultimately, the regression tree provides predictions for the target variable based on the specific path followed through the tree. The predicted value is typically the average or the majority value of the target variable in the leaf node where a new data point falls. This approach allows for the prediction of continuous values, making regression trees suitable for regression-type problems such as price estimation.

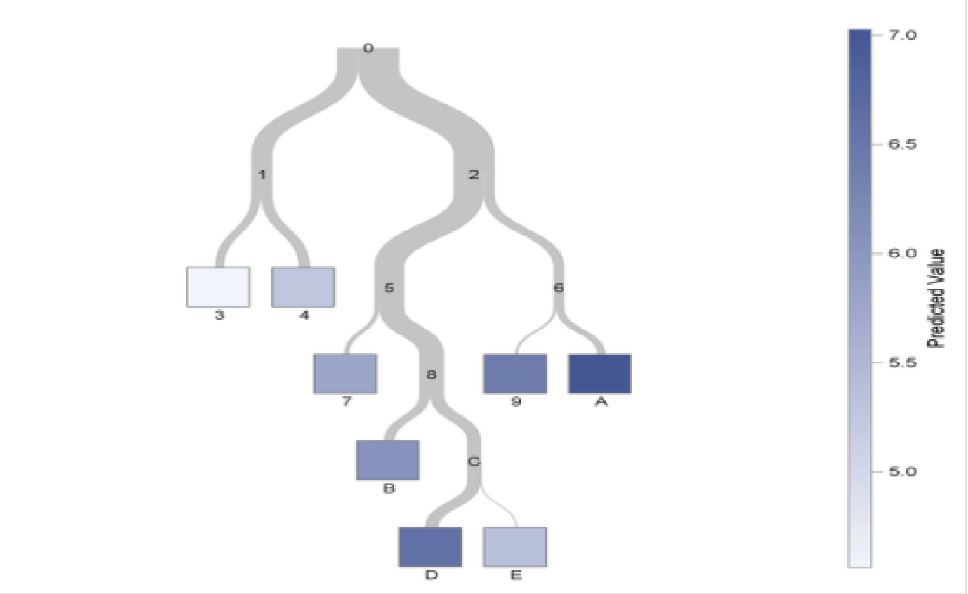

Fig.2.4. regression for log salary.

3.3.3. When to use Classification and Regression:

Classification trees are specifically used when the goal is to split a dataset into different classes or categories based on the response variable. Typically, these classes are binary, such as "Yes" or "No," and they are mutually exclusive. However, there are cases where the response variable has more than two classes, and in such instances, variant algorithms of classification trees, such as multiclass classification trees, are utilized.

On the other hand, regression are employed when the response variable is continuous and the objective is to predict a numerical value. Examples of response variables suitable for regression include property prices, temperature readings, or any other continuous variable. Regression are specifically designed for prediction-type problems where the focus is on estimating numerical values rather than classifying into discrete categories.

Therefore, classification trees are primarily used for classification-type problems, where the goal is to assign data points to specific classes, while regression trees are used for prediction-type problems, where the aim is to forecast continuous values [9].

3.3.4. How Classification and Regression Work:

In a classification tree, the dataset is split based on the homogeneity or purity of the data. For example, if we have variables like income and age to predict whether a consumer will buy a particular phone, the data is divided based on the patterns observed in the training data. If, for instance, 95% of people older than 30 have bought the phone, the data is split at that point, and age becomes the top node in the tree. This split results in a "95% pure" data subset. Measures of impurity such as entropy or Gini index are used to quantify the homogeneity of the data in classification trees.

In a regression tree, a regression model is applied to the target variable using each of the independent variables. The data is then split at multiple points for each independent variable. At each split point, the squared error between the predicted values and the actual values is calculated, resulting in a "Sum of Squared Errors" (SSE). The SSE is compared across the variables, and the variable or split point with the lowest SSE is chosen. This process is repeated recursively, creating a tree structure that optimizes the prediction of the continuous target variable.

Overall, both classification and regression trees use recursive splitting based on specific criteria to create a tree-like structure that represents the patterns and relationships within the data. The goal is to find the best splits that maximize homogeneity or minimize error, allowing for accurate predictions or classification [10].

4. Main metrics of ML models evaluation:

The process of evaluating models goes through a set of different criteria:

Choosing the right way to evaluate a classification model is as important as choosing the classification model itself, if not more. Sometimes, accuracy might not be the best way to evaluate how a classification model performs.

In this section, we will present and explain the different metrics used to evaluate results for classification models.

- Accuracy:

Accuracy is the conventional method of evaluating classification models. Accuracy is defined as the proportion of correctly classified examples over the whole set of examples.

Accuracy = (Number of correct predictions) / (Overall number of predictions)

Accuracy is very easy to interpret, in practice, it is only used when the dataset permits it. It is not completely unreliable as a method of evaluation, but there are other, and sometimes better, methods that are often overlooked.

When you only use accuracy to evaluate a model, you usually run into problems. One of which is evaluating models on imbalanced datasets.

- Confusion Matrix:

A confusion matrix is an error matrix. It is presented as a table in which the predicted class is compared with the actual class. Understanding confusion matrices is of paramount importance for understanding classification metrics, such as recall and precision.

When performing classification predictions, there's four types of outcomes that could occur.

a- True positives:

The expectation of the observation's membership in a class, and its validity in that membership.

- **True negatives:**

Observation does not belong to a class and it is true that it belongs to the same class.

c- False positives:

The expectation of an observation not belonging to a class, when in fact it does.

d- False negatives:

The expectation that an observation does not belong to a category when in fact it is.

These four outcomes are often plotted on a confusion matrix. The following confusion matrix is an example for the case of binary classification.

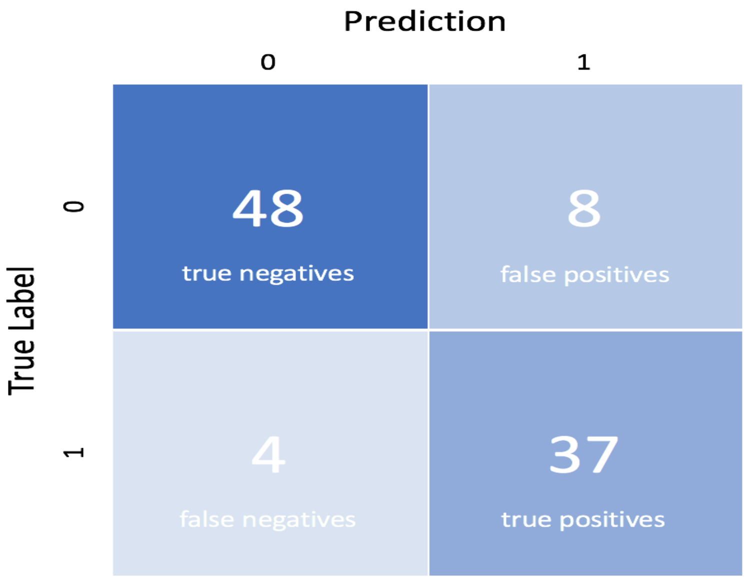

Fig.2.5. Confusion matrix heat map

- Precision (Positive Predicted Value):

Precision is defined as the number of true positives divided by the sum of true and false positives. Precision expresses the proportion of data correctly predicted as positive. Using it as a metric, you can define the percent of the predicted class inside the data you classified as that class. In other words, precision helps you measure how often you correctly predicted that a data point belongs to the class your model assigned it to. The equation for it is:

Precision = (True Positive) / (True Positive + False Positive)

- Recall (Sensitivity, True Positive Rate):

The recall is defined as the number of true positives divided by the sum of true positives and false negatives. It expresses the ability to find all relevant instances in a dataset. Recall measures how good your model is at correctly predicting positive cases. It’s the proportion of actual positive cases which were correctly identified. The equation for recall is:

Recall = (True Positive) / (True Positive + False Negative)

- F1 score:

A common metric that combines validation and recovery criteria to provide a balanced estimate of model performance. The F1 criterion is calculated as the geometric mean between Precision and Recall, and is useful in cases where the data is imbalanced.

- Diagnostic Curve: (Receiver Operating Characteristic - ROC):

This measure is used to evaluate the performance of the model when there is a balance between recovery and false positive. It is based on drawing a curve showing the model's ability to distinguish between different categories. Performance is measured by the Area Under the Curve (AUC), where the greater the AUC value, the better the performance[SW8].

5. Related works

Besides the rise development of communication technologies and mobile networks, USSD prepaid cards remain the most commonly used form to credit/recharge mobile phone accounts in Algeria. However, checking fraud in such operation is a significant problem that causes hundreds of millions of dollars in losses annually. In this section, we present some related works that delved into the subject of fraud and some proposed solutions.

In [11], the study collected information on the history of checks, forms of check fraud, victimization, and methods for prevention and detection. While artificial intelligence has made many recent advancements in machine learning and computer vision, financial institutions have not kept pace with anti-fraud measures.

The research concludes that financial institutions must take a modern approach to check fraud by incorporating machine learning into real-time reviews. This is necessary to adequately protect victims and strike a balance between educating customers, complying with regulations, and providing valuable and fast alerts. By doing so, financial institutions can better combat check fraud and ensure that vulnerable consumers are not disproportionately affected.

In her thesis [12] author presents that many businesses believe that in-depth knowledge of either machine learning or statistical methods is enough to be a predictive modeler. Then, she aims to dispel the fallacy and demonstrate that the mathematics behind the model is just as important, if not more important, than the computer science needed to implement it.

The thesis explores existing methodologies for fraud detection proposed by academic professionals worldwide and evaluates their accuracy, efficiency, and reliability on a large dataset downloaded from a public website. The methods analyzed are hidden Markov models (HMM), convolutional neural networks (CNN), and support vector machines (SVM). For each method, the thesis presents the history and motivation, theoretical framework, strengths and weaknesses, and numerical examples performed in either SAS Enterprise Miner or R.

As credit card fraud is a significant and growing problem in the financial market, researchers have focused on detecting fraudulent behavior early using advanced machine learning techniques. In his thesis [13], various predictive models, such as logistic regression, random forest, and XGBoost, in combination with different resampling techniques, have been applied to predict fraudulent transactions.

The experimental results showed that random forest in combination with a hybrid resampling approach of Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links removal performed better than other models.

6. Conclusion:

In this chapter, we have briefly introduced machine learning with its different types and algorithms. Then we briefly explained the principle of classification and regression, and then we explained the main criteria for evaluating the machine learning model.

Chapter 3:

Proposed Approach

Implementation & Results

1. Introduction:

In this chapter, we explain the data and algorithms that we have done in this study to calculate the accuracy of the data each time with several algorithms, including machine learning and deep learning algorithms, with the aim of analyzing the weaknesses of that data and finding out how difficult it is to encode it.

2.Proposed Approach:

The fundamental stages of the machine learning process consist of data collection and preparation, model selection, model training using training data, performance evaluation of the model, and visualization of the obtained results.

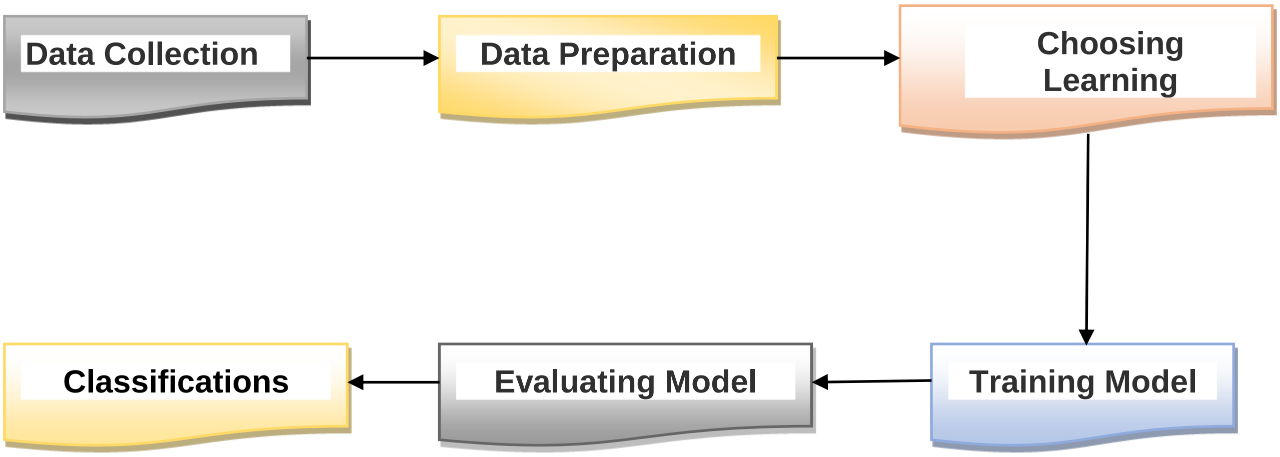

Fig.3.1. Machine Learning Workflow

3. Detailed Steps:

3.1. Data Collection:

For us, collecting data on used shipping labels was among the most difficult, given the scarcity of available resources. Many have resorted to charging their balance through the Internet or the services of telecom companies, which leads to the unavailability of paper cards that contain the required data. However, we did everything we could to collect as many used cards as possible, by visiting our neighboring shops and various communication agencies and asking for their cooperation in collecting these cards.

3.2. Data processing:

In processing and calibrating the data, we used several well-known approaches to ensure comprehensive coverage of the data. Specifically, we used multiple algorithms such as Luhn's algorithm, barcode, and Upc to accommodate a wide range of possibilities in the data. These methods have been applied to enhance accuracy and reliability in our data processing pipeline.

We also made additional divisions of the Luhn and normal data, so that it was divided into 3 forms, so that the first form we considered the numbers as one entity, the second form was divided into 4 parts, and the third form was divided into 15 digital (each number alone)

- Luhn's Algorithm: The Luhn's Algorithm (also known as Mod 10) is a simple checksum formula used as a first line of defense on e-commerce and government websites to verify the validity of a variety of identification numbers, such as IMEI card numbers, US National and Canadian Social Security numbers, and others.

- Universal Product Algorithm: A Universal Product Code (UPC) is a code found on a product's packaging that helps identify the product. 12 digits below it. UPC barcodes are scannable codes that are designed to make it easier to identify product characteristics, such as brand, size, designation, or color. The last digit of the UPC code is also associated with the verification number.

- Codabar algorithm: its plan is designed when printing on dot-matrix and typewriter printers.

We have this example:

Type Data	Format 1	Format 2	Format 3
Data Normal	60720602504590	6072-2060-2504-590	6-0-7-2-2-0-6-0-2-5-0-4-5-9-0
Data Luhn	030520602108590	0305-2060-2108-590	0-3-0-5-2-0-6-0-2-1-0-8-5-9-0
Data Upc	0-6-0-7-6-0-18-0-6-5-0-4-15-9-0		
Data Codabar	0-6-0-7-4-0-12-0-4-5-0-4-10-9-2-0		

Tab.3.1. Data conversion example

3.3. Choosing Learning Algorithm:

After selecting a set of algorithms to process the collected data for an analytical study of recharge cards, we proceeded to evaluate their performance by calculating the accuracy.

By calculating the accuracy for each algorithm, we can gain insights into their effectiveness in predicting recharge card patterns. This information is crucial for identifying the most accurate algorithm and making informed decisions on which algorithm to use for further analysis or deployment in practical scenarios.

Among the selected algorithms for the analytical study of recharge cards, we have chosen K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Decision Tree (DT), and Random Forest (RF).

3.4. Model Selection and Training:

After completing the data processing stage, the next step is to divide the data set into two parts. the training data and the test data at 80% and 20% respectively. Then, machine learning and deep learning algorithms are applied to the data.

3.5. Evaluating Model:

At this stage, he proposed a set of algorithms to calculate the accuracy of the data in each type of data, so that we took machine learning and deep learning algorithms from them, to see which one is more correct to calculate the accuracy.

4. Results analysis:

We have utilized a serie of algorithms to process a diverse range of data sets, employing various forms of analysis. The outcomes obtained from these analyses have been documented and will be presented.

Type Dataset	Number blocks	algorithm name			
		ANN	RF	DT	KNN
Normal	1_BLOCK	61.51%	41.85%	41.57%	56.46%
	4_BLOCKS	63.20%	54.49%	47.47%	52.52%
	15_ digital	60.39%	54.21%	48.03%	51.40%
Luhn	1_BLOCK	61.51%	40.73%	40.44%	50.56%
	4_BLOCKS	54.49%	50.56%	48.03%	48.31%
	15_ DIGITAL	60.95%	52.52%	47.47%	47.19%
Codabar	16_ DIGITAL	60.67%	54.21%	48.31%	52.52%
UPC	15_ DIGITAL	59.26%	55.05%	49.43%	46.91%

Tab.3.2. accuracy results

- Accuracy of the model for Ann data (Codabar data):

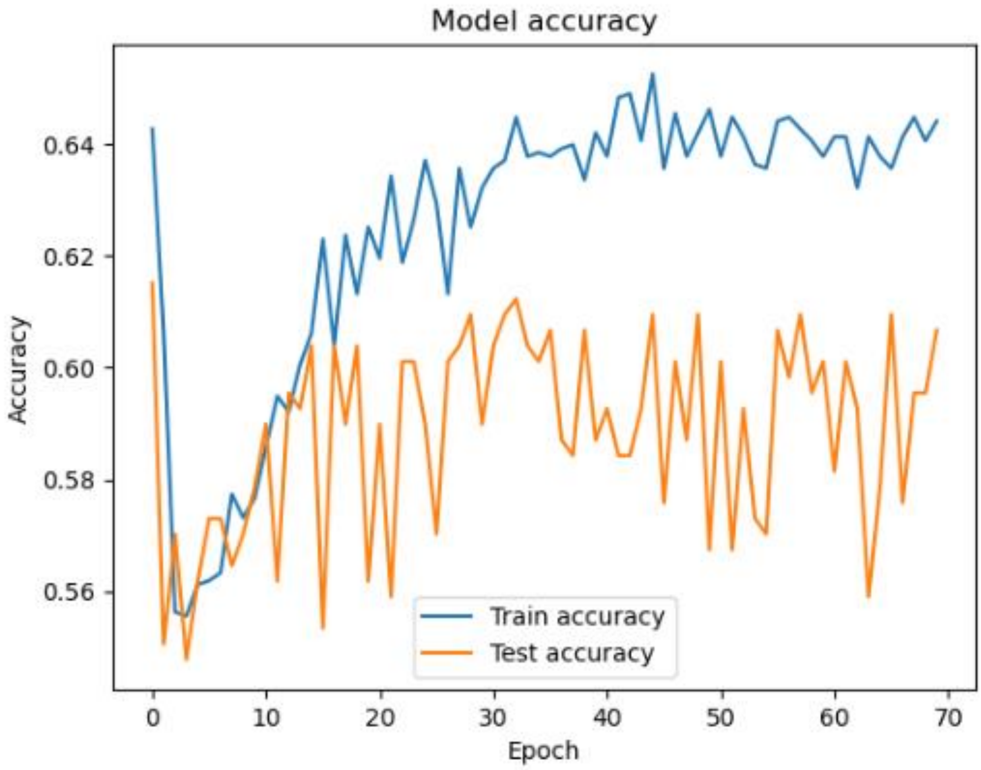

Fig.3.2. Model accuracy

- Accuracy of the model for DT data(Codabar data):

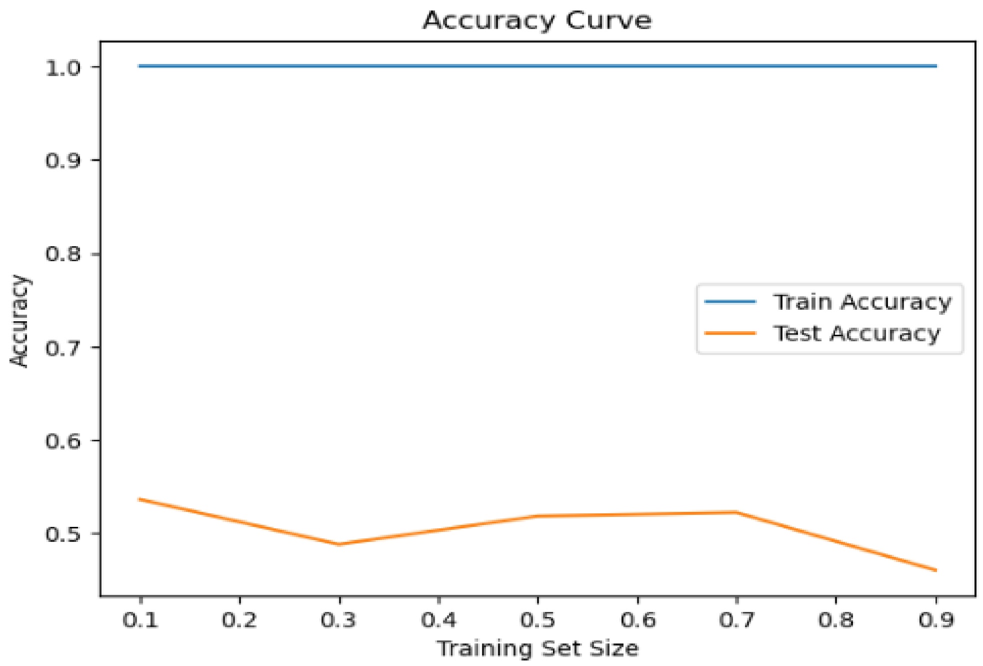

Fig.3.3. Model accuracy

- Accuracy of the model for KNN data (Codabar data):

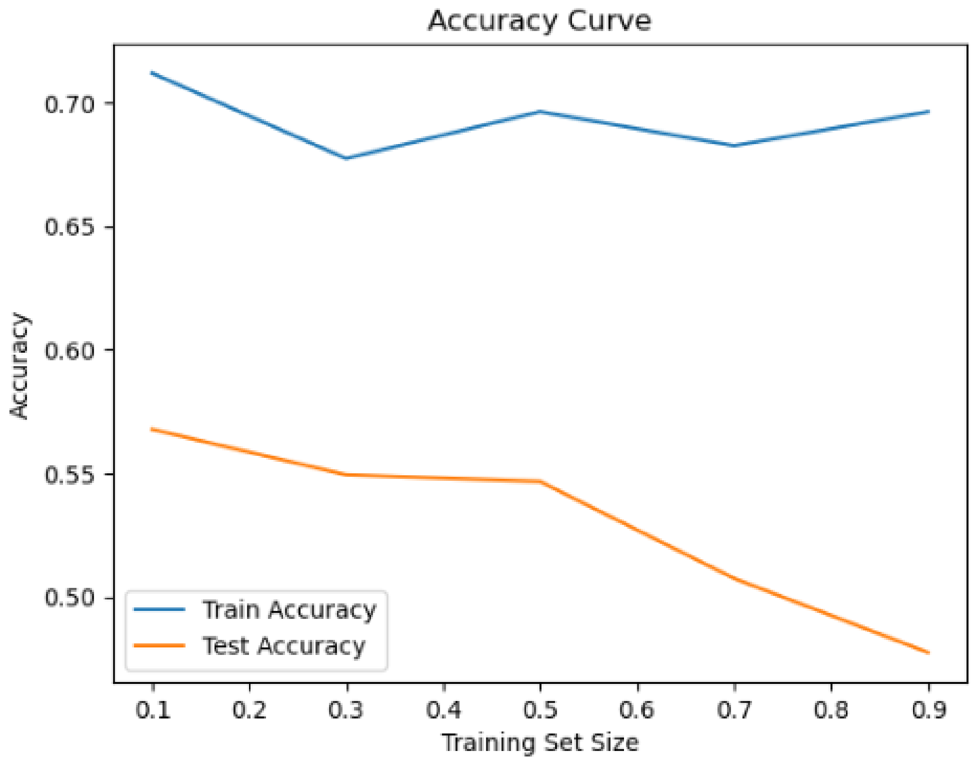

Fig.3.4 Model accuracy

- Accuracy of the model for RF data (Codabar data):

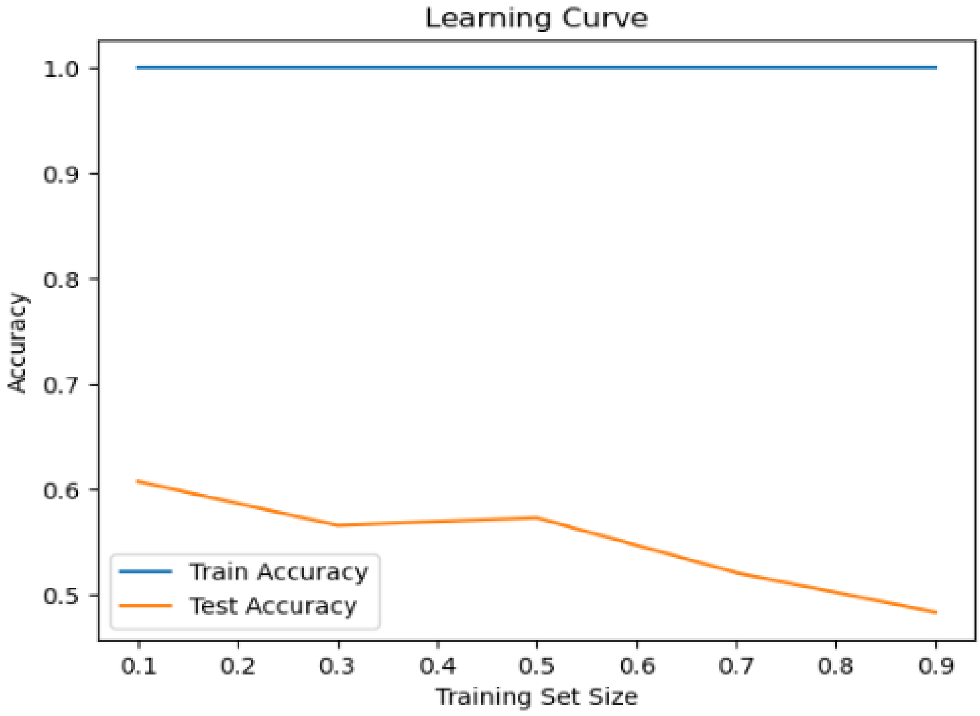

Fig.3.5 Model accuracy

5.Work Tools

5.1 .Jupyter Notebook:

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

5.2 .Python:

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

Results and discussion:

In our thesis, we have used three of the most known algorithm for validating code such as Luhn algorithm used for validating credit cards and IMEI, and Algorithms for Codabar validation and UPC. In addition, we have analyzed the USSD recharge cards digits with three possibilities (as one block, four blocks and as 15 digits) to give the machine learning more facility to detect patterns if exists.

As result, we have concluded that the best algorithm was the KNN and the better accuracy was for the case of analyzing with 15 digits.

Finally, we have concluded that analyzing USSD recharge codes using machine learning techniques can provide useful insights into codes patterns. However, the studied mobile phone operator uses a good algorithm for generating USSD recharge code. Moreover, we advocate to enhance such algorithms by including the validity period, and the amount of card to better strength that algorithms.

General conclusion

In conclusion, the use of machine learning techniques to analyze USSD recharge codes showed a significant level of accuracy (63.20%). In fact, this approach provides quick and more accurate analysis of USSD recharge codes, which can help telecommunication operators to better improving their algorithms or techniques used for USSD recharge codes generation. However, it is important to note that accuracy depends highly on the quality and volume of data used to train the model, as well as the complexity of the AI models used. Therefore, it is recommended to continuously gathering additional data, improve their quality and explore new AI techniques to further improve the accuracy of results.

.

References

[1] :Heindl, E. (2009). Mobile Network. E-Business Technology, Shirin Faghihi, (232493).

[2] :Eberspächer, J., Bettstetter, C., Vögel, H. J., & Hartmann, C. (2008). GSM-architecture, protocols and services. John Wiley & Sons.

[3] :Le Bodic, G. (2005). Mobile messaging technologies and services: SMS, EMS and MMS. John Wiley & Sons.

[4] :USSD Interworking Guidelines Version 1.0, Official Document IR.45. GSM Association. July 2013. Available at :www.gsma.com/newsroom/wp-content/uploads/2013/07/IR.45-v1.0.pdf

[5] :Mekala, S. R. (2015). Mobile Credit Using Gsm Network: TopupFor Mobile Phones. Master thesis, Faculty of Computing Blekinge, Institute of Technology SE-371 79, Sweden.

[6] :Training Guide Voucher Processing. (2009). Columbia University's FinancialSystems. Available at:https://www.finance.columbia.edu/sites/default/files/content/Training/Documents/Voucher_Processing_TRAIN.pdf

[7] :Hussein, K. W., Sani, N. F. M., Mahmod, R., & Abdullah, M. T. (2013). Enhance Luhn algorithm for validation of credit cards numbers. Int. J. Comput. Sci. Mob. Comput, 2(7), 262-272.

[8]: A total guide to barcoding (2016). GSM Barcoding. Available at :https://www.barcoding.co.uk/wp-content/uploads/2016/02/E-Book_Total-Guide-to-Barcoding.pdf

[9]: DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY (Autonomous Institution – UGC, Govt. of India). MACHINE LEARNING [R17A0534] LECTURE NOTES

[10]: Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020, May). Credit card fraud detection using machine learning. In 2020 4th international conference on intelligent computing and control systems (ICICCS) (pp. 1264-1270). IEEE.

[11] :Rose, L. M. (2018). Modernizing check fraud detection with machine learning (Doctoral dissertation, Utica College).

[12]: Sarah E. W. (2017). Modernizing check fraud detection with machine learning (Master of Science in Applied Statistics dissertation, California State University).

[13]: Shakya, R. (2018). Application of machine learning techniques in credit card fraud detection (Doctoral dissertation, University of Nevada, Las Vegas).

Sites web :all links have been checked in 27/05/2023

[SW1] :https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[SW2] :https://www.aps.dz/ar/sante-science-technologie/130322-47-2022

[SW3] :https://shorturl.at/elDNV (Tadamsanews)

[SW4] :https://www.spiceworks.com/tech/networking/articles/what-is-gsm/

[SW5]:https://data.worldbank.org/indicator/IT.CEL.SETS?end=2021&locations=DZ &start=1960

[SW6] :https://blog.ussd.directory/how-ussd-works-an-explainer/

[SW7] :https://tritonstore.com.au/what-is-codabar-barcode/

[SW8]: https://www.edlitera.com/blog/posts/evaluating-classification-models