People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

UNIVERSITY OF
ECHAHID HAMA LAKHDER EL OUED
FACULTY OF EXACT SCIENCES
Computer Science department

End of study memory

Presented for the Diploma of

ACADEMIC MASTER

Domain : **Mathematics and Computer Science**

Option : **Computer Science**

Speciality: **Distributed Systems and Artificial Intelligence**

Presented by: Bousbia Laiche Meriem

Tedjani Mabrouka

Theme

Machine learning based system for fetus gender prediction

.

Defended on *15-06-2022* in front of the jury composed of:

Mr. Zaïz Faouzi	El-Oued university	President
Ms. Belila Khaoula	El-Oued university	Examinator
Mr. Boucherit Ammar	El-Oued university	Supervisor

Academic year 2021/2022

Acknowledgments

One of the secrets of success in life is looking to the past and rewarding ourselves for the achievements we have achieved with praise and thanks to God, optimism and spreading positive energy.

To those who supported my faltering mistakes, my thanks and appreciation are the symbols of giving and loving my mother and my father.

My teacher, for every creator an achievement, for every thank you is a poem, for every position there is an article, and for every success, thanks and appreciation. Many thanks to the teacher, supervisor and mentor "Ammar Boucherit".

We extend our thanks and appreciation to the jury for their evaluation of this positioned work.

We came on this day to reap the fruits of the past years and take the first step in my path. My goals I want to present bouquets of flowers to everyone who was the reason for this work. We thank all of our family, friends and teachers.

Thanks….

M. Tedjani
M. Bousbia Laiche

Dedicate

If I could gift time, I would have gifted mine to yours, and if my heart was an open book, you would find your names on the first lines, all words mean nothing in your presence. To the ones who lit the first candle of my youth to my adulthood, to my shelter and haven, to my my beloved father "Mohammed", and to the human that gives without waiting for a return, my dear mother "Massaouda".

Our grand parents are above us, and a sincere godly blessing on them, to my grand father "Ali" and my grandmother "Sahra".

To my sisters, the dear "Basma", the beloved "Israa", the kind "Imane", and the pretty "Nour El-Houda", and the precious "Nardjesse". Who have been all of immense help.

To all who helped in making this dissertation better. To both teachers, "Miss Aicha" and "Mr Brahim", and to my workmate "Meriem", and to all my college friends, to the entire class of 2022.

Mabrouka Tedjani

Dedicate

I would like to dedicate this humble act to everyone who has encouraged me throughout this work.

To those who have no equal in the universe, whom God has commanded us to honor, to those who have done so much, and who have paid the irretrievable, here are these words, my dear mother and father, I dedicate this research to you. You have been my best support throughout my college career.

To my dear brothers and sisters for their advice and the support.

To my companion Mabrouka for her availability, understanding and assistance during the completion of this work.

To the travel companion, and to the daily friend in its sweetness and bitterness: my dear husband, I dedicate this research to you as an expression of my thanks for your continued support.

I must not forget my teachers who had the greatest role in supporting me and providing me with valuable information. . . .

Everyone who helped me directly or indirectly in the implementation of this project.

Everyone who knows me.

Thanks.

Meriem Bousbia Laiche

Abstract

The desire to reveal the gender of the fetus is not new, but a curiosity as old as humanity. For that, many assumptions have arisen among ancestors since medicine had not yet reached a relevant scientific answer to the hypothesis of determining the gender of the fetus at very advanced stages of pregnancy like the hypotheses of the Chinese and Mayans.

In this context, we have performed a retrospective database analysis of more than 300,000 birth records; in which we aim to use and compare some supervised machine learning classification algorithms; to find the best one in answering this ancient question according to some characteristics such as according to the maternal age during pregnancy and the month of conception, then study the possibility of comparing the results with some of the older hypotheses in the field.

The obtained results are very encouraging for scientific research to consider such features. At the same time, we aim to extend our study and build a deep learning based model for training larger datasets with more features.

Keywords: machine learning, fetus gender prediction

خلاصة

إنّ الرغبة في الكشف عن جنس الجنين ليست جديدة ، لكنها فضول قديم قدم البشرية. لذلك ، نشأت العديد من الافتراضات بين الأجداد لأن الطب لم يتوصل بعد إلى إجابة علمية رصينة لفرضية تحديد جنس الجنين في مراحل متقدمة جدا من الحمل مثل فرضيات الصينين والمايويين.
في هذا السياق ، أجرينا تحليلًا تراجعيا لقاعدة البيانات تحوي أكثر من ٣٠٠٠٠٠ سجل ولادة؛ التي نهدف فيها إلى استخدام ومقارنة بعض خوارزميات تصنيف التعلم الآلي الخاضعة للإشراف ؛ للعثور على الأفضل في الإجابة على هذا السؤال القديم وفقًا لبعض الخصائص مثل عمر الأم أثناء الحمل وشهر الحمل ، ثم دراسة إمكانية مقارنة النتائج ببعض الفرضيات القديمة في هذا المجال.
النتائج التي تم الحصول عليها مشجعة للغاية للبحث العلمي للنظر في هذه الميزات. في الوقت نفسه ، نهدف إلى توسيع دراستنا وبناء نموذج قائم على التعلم العميق لتدريب مجموعات بيانات أكبر بمزيد من الميزات.

الكلمات االمفتاحية
التعلم الآلي، التبؤ بجنس الجنين.

Résumé

Le désir de révéler le genre du fœtus n'est pas nouveau, mais une curiosité aussi ancienne que l'humanité. Pour cela, de nombreuses hypothèses ont surgi chez les ancêtres puisque la médecine n'avait pas encore abouti à une réponse scientifique pertinente à l'hypothèse de déterminer le genre du fœtus à des stades très avancés de la grossesse comme les hypothèses des Chinois et des Mayas.

Dans ce contexte, nous avons effectué une analyse rétrospective d'une banque de données de plus de 300 000 actes de naissance ; dans lequel nous visons à utiliser et à comparer certains algorithmes de classification d'apprentissage automatique supervisé ; afin de trouver le meilleur pour répondre à cette question ancienne selon certaines caractéristiques telles que selon l'âge maternel pendant la grossesse et le mois de conception, puis étudier la possibilité de comparer les résultats avec certaines des hypothèses les plus anciennes dans ce domaine.

Les résultats obtenus sont très encourageants pour que la recherche scientifique considère de telles caractéristiques. Dans le même temps, nous visons à étendre notre étude et à construire un modèle basé sur l'apprentissage en profondeur pour former des ensembles de données plus volumineux avec plus de fonctionnalités.

Mots clés: apprentissage automatique, Prédiction du genre du fœtus.

Table of Contents

List of Figures

List of Tables

General Introduction

Although the idea of gender selection may be controversial, it has been part of human's history from time immemorial. Therefore, out of fun and curiosity, many couples —with their different cultural levels— attempted to apply some old thoughts and opinions on this subject. Precisely, couples have attempted gender preselection of their future child by using alkalinized or acidified liquids or determining ovulation days to boost the likelihood of a female or male baby. In addition, popular (folk) methods include some beliefs related to wind direction, precipitation, temperature, or the phases of the moon at the time of conception. There is also an old popular belief that burying the placenta of the current newborn under the nut tree will ensure that a male child will be born the next time [4].

On the other hand, machine learning is a new scientific and practical pattern that has recently become popular in a large number of scientific, medical, industrial and other fields, especially with the rapid developments that the world is witnessing with the revolution of big data, which provided an important source for information exploration (Data mining) within data that we did not previously think was of great importance.

In fact, machine learning is one of the branches of artificial intelligence, which bases its work on data and algorithms, to imitate the way humans learn, and gradually improve it. Indeed, some machine learning and artificial intelligence experts see machine learning as the next real challenge for researchers since it has cognitive abilities to infer, classify, predict and solve problems that rival the skills of human researchers [5]. For instance, Google has developed a machine learning algorithm to help identify cancerous tumors on mammograms. . . etc.

In this context, gender and age prediction -as a research axis- have received a lot of attention in recent years since it is widely applied in judicial investigations to determine the identity of terrorists or criminals and others in acts of theft, violence or traffic accidents ...etc.

For that, different factors have been proposed and tested to determine a someone gender through Facial images [6], Voice [7], Finger print [8], etc. Consequently, there are a variety of online web applications and tools that are capable of pointing out the gender from user's characteristics [9, 10, 11].

1. Problem context

The ability to anticipate the gender of a future child is a topic of great public interest. Although the likelihood of having a child of either gender is close to 50% for most cases, there are a number of factors that can significantly alter this value. Although some research conducted by our ancestors revealed that previous pregnancies, maternal age, and the month of conception can influence gender determination, this hypothesis was not widely recognized among clinicians because it was not scientifically approved.

2. Scope of the thesis

In the present work, we will focus on the gender prediction of newborn (fetus) by performing a retrospective dataset analysis of more than 300.000 birth records collected from a number of municipalities in El-Oued State, and using some supervised machine learning classification algorithms such as Random Forest, KNN and others in order to find the best one in answering this ancient question. In addition, we also aim to compare the obtained results with the calculated accuracy of both Chinese and Mayan hypotheses in this field since they are based on the same used features like our work.
Unfortunately, we hoped to collect a number of additional features for our study, such as the blood group of the parents, etc., and although the intervention of the dean of the faculty of Exact Sciences at El-Oued University with the director of El-Oued Education Directorate, we were unable to do so due to a number of obstacles.

3. Thesis organization

This thesis is composed of three chapters. The first one presents the main related concepts to machine learning. In the second chapter, we present some previous works on the subject of baby gender prediction. The third one gives a detailed overview about our work and obtained results. Finally, according to the obtained results, this thesis closes with a discussion of possible future research directions.

Chapter 1

Machine learning

1.1: Introduction

Today, the world has witnessed a strong entry of information technology, especially machine learning. So we wanted to talk about it in this chapter about its importance and the rate of spread of its applications in our daily lives, such as Facebook, and speech recognition, in order to predict diseases of the day, and many more. and to identify its basic types that are supervised, unsupervised, and reinforcement. and also criteria that allow us to choose the right kind of approach to arrive at an ideal model to achieve the desired objective of the study.
Machine learning is a sub-field of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people [12].

1.2: Machine learning

1.2.1: Definition

Machine learning is a data analysis method that automates the construction of analytical models. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [13].
Automatic Learning or Machine Learning, is to use algorithms so that they can say something interesting based on a set of data without having to write any specific code for the problem [14].

1.2.2: Why ML is needed ?

Machine learning technology is now necessarily new; machine learning algorithms have existed for years, but machine learning processes have recently taken prominence due to several important technological improvements, including:

- Wider access to large volumes and varieties of data, especially the development and ubiquity of “big data”.

- Much more affordable data storage solutions, which helped make big data sets available to more organizations and for a much wider variety of applications.
- Increasing processing power that allows computers and specifically AI applications to complete calculations much faster than ever before

These developments set the scene for machine learning to produce much better results than it was historically capable of providing, allowing machine learning applications to provide value to virtually every industry and business activity.

1.2.3: Application of ML

1. **Virtual Personal Assistants**
   Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice. All you need to do is activate them and ask "What is my schedule for today?", "What are the flights from Germany to London", or similar questions. For answering, your personal assistant looks out for the information, recalls your related queries, or send a command to other resources (like phone apps) to collect info. You can even instruct assistants for certain tasks like "Set an alarm for 6 AM next morning", "Remind me to visit Visa Office day after tomorrow".
   Machine learning is an important part of these personal assistants as they collect and refine the information on the basis of your previous involvement with them. Later, this set of data is utilized to render results that are tailored to your preferences [15].
   Virtual Assistants are integrated to a variety of platforms. For example:
   - **Smart Speakers:** Amazon Echo and Google Home.
   - **Smartphones:** Samsung Bixby on Samsung S8.
   - **Mobile Apps:** Google Allo.
2. **Predictions while Commuting**
   Traffic Predictions: We all have been using GPS navigation services. While we do that, our current locations and velocities are being saved at a central server for managing traffic. This data is then used to build a map of current traffic. While this helps in preventing the traffic and does congestion analysis, the underlying problem is that there are less number of cars that are equipped with GPS. Machine learning in such scenarios helps to estimate the regions where congestion can be found on the basis of daily experiences.
   Online Transportation Networks: When booking a cab, the app estimates the price of the ride. When sharing these services, how do they minimize the detours? The answer is machine learning. Jeff Schneider, the engineering lead at Uber ATC reveals in a an interview that they use ML to define price surge hours by predicting the rider demand. In the entire cycle of the services, ML is playing a major role [16].

3. **Videos Surveillance**

   Imagine a single person monitoring multiple video cameras! Certainly, a difficult job to do and boring as well. This is why the idea of training computers to do this job makes sense.

   The video surveillance system nowadays are powered by AI that makes it possible to detect crime before they happen. They track unusual behaviour of people like standing motionless for a long time, stumbling, or napping on benches etc. The system can thus give an alert to human attendants, which can ultimately help to avoid mishaps. And when such activities are reported and counted to be true, they help to improve the surveillance services. This happens with machine learning doing its job at the backend.

4. **Social Media Services**

   From personalizing your news feed to better ads targeting, social media platforms are utilizing machine learning for their own and user benefits. Here are a few examples that you must be noticing, using, and loving in your social media accounts, without realizing that these wonderful features are nothing but the applications of ML.

   - **People You May Know:**Machine learning works on a simple concept: understanding with experiences. Facebook continuously notices the friends that you connect with, the profiles that you visit very often, your interests, workplace, or a group that you share with someone etc. On the basis of continuous learning, a list of Facebook users are suggested that you can become friends with.
   - **Face Recognition:**You upload a picture of you with a friend and Facebook instantly recognizes that friend. Facebook checks the poses and projections in the picture, notice the unique features, and then match them with the people in your friend list. The entire process at the backend is complicated and takes care of the precision factor but seems to be a simple application of ML at the front end.
   - **Similar Pins:** Machine learning is the core element of Computer Vision, which is a technique to extract useful information from images and videos. Pinterest uses computer vision to identify the objects (or pins) in the images and recommend similar pins accordingly.

5. **Email Spam and Malware Filtering**

   - There are a number of spam filtering approaches that email clients use. To ascertain that these spam filters are continuously updated, they are powered by machine learning. When rule-based spam filtering is done, it fails to track the latest tricks adopted by spammers. Multi Layer Perceptron, C 4.5 Decision Tree Induction are some of the spam filtering techniques that are powered by ML [17].
   - Over 325, 000 malwares are detected everyday and each piece of code is 90–98% similar to its previous versions. The system security programs that are powered by machine learning understand the coding pattern. Therefore, they detects new malware with 2–10% variation easily and offer protection against them.

6. **Online Customer Support**

   A number of websites nowadays offer the option to chat with customer support representative while they are navigating within the site. However, not every website has a live executive to answer your queries. In most of the cases, you talk to a chatbot. These bots tend to extract information from the website and present it to the customers. Meanwhile, the chatbots advances with time. They tend to understand the user queries better and serve them with better answers, which is possible due to its machine learning algorithms.

7. **Search Engine Result Refining**

   Google and other search engines use machine learning to improve the search results for you. Every time you execute a search, the algorithms at the backend keep a watch at how you respond to the results. If you open the top results and stay on the web page for long, the search engine assumes that the the results it displayed were in accordance to the query. Similarly, if you reach the second or third page of the search results but do not open any of the results, the search engine estimates that the results served did not match requirement. This way, the algorithms working at the backend improve the search results.

8. **Product Recommendations**

   You shopped for a product online few days back and then you keep receiving emails for shopping suggestions. If not this, then you might have noticed that the shopping website or the app recommends you some items that somehow matches with your taste. Certainly, this refines the shopping experience but did you know that it's machine learning doing the magic for you? On the basis of your behaviour with the website/app, past purchases, items liked or added to cart, brand preferences etc.the product recommendations are made.

9. **Online Fraud Detection**

   Machine learning is proving its potential to make cyberspace a secure place and tracking monetary frauds online is one of its examples. For example: Paypal is using ML for protection against money laundering. The company uses a set of tools that helps them to compare millions of transactions taking place and distinguish between legitimate or illegitimate transactions taking place between the buyers and sellers [18].

1.2.4: Difference between Traditional IA and ML

In traditional programming, all the rules that the program needs to provide a result are written. Then, the program is given input data that will be processed based on the rules and in this way the expected results are obtained.
On the contrary, in Machine Learning the system is trained instead of being programmed explicitly. For this, the input data and the expected results are needed. In the end we get a set of rules, also called a model [19].

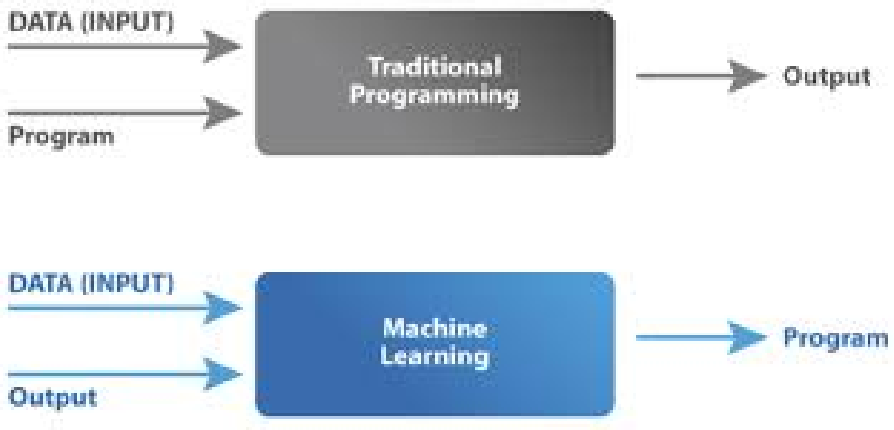

Figure 1.1: Machine Learning vs Traditional Programming[1].

1.3: Main types of learning

In this section we will talk about the 3 most important types of learning in machine learning and about the most important algorithms used in the project.

1.3.1: Supervised learning

Supervised learning is one that works with labeled data. This type of learning tries to find a way to recognize similarities between data with the same label in order to be able to differentiate between data with different labels, and in this way, manage to classify data that do not have a label by assigning them one of the different categories. These training data have the following structure: given some data (x1,1), . . . , (xn, yn) where x= x1, x2, . . . , xD , N is the number of examples, X represents the characteristics with D being the dimensions and finally Y represents the labels. A learning algorithm tries to obtain a function G: X→ Y, where X is the input and Y is the output .
In this type of learning, 2 categories are differentiated according to the type of label: classification and regression.

3.1.1 Classification

In this category of supervised learning, the algorithm focuses on classifying the data into different specific categories. The algorithms try to define to which category the input data to be classified belongs to according to their characteristics.
It is often used for cases such as defining the category of a news item, knowing if a message is spam or not, etc.

Types of ML Classification Algorithms

Classification Algorithms can be further divided into the Mainly two category:

- **Linear Models:**
  - Logistic Regression
  - Support Vector Machines. .
- **Non-linear Models:**
  - K-Nearest Neighbours.
  - Kernel SVM.
  - Naïve Bayes.
  - Decision Tree Classification.
  - Random Forest Classification.

Classification Algorithms in Machine Learning

1. **logistic regression**

   Logistic Regression is a classification algorithm used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable containing data coded as 1 – 0, yes – no, open – closed, etc.
   This binary logistic model is used to estimate the probability of a binary response based on one or more independent or predictor variables. Lets say that the presence of a risk factor increases the probability of a given result a specific percentage.
   logistic regression is a predictive analysis. Logistic regression requires fairly large sample sizes.
   The reason why logistic regression is widely used, despite advanced algorithms like deep neural networks, is because it is very efficient and does not require too many computational resources which makes it affordable to run in production.

2. **K-Nearest Neighbors**

   The KNN algorithm is one of the simplest ranking algorithms, even with such simplicity it can give highly competitive results. It belongs to the supervised learning domain and can be used for pattern recognition, data mining, and intrusion detection.
   It is a robust and versatile classifier that is often used as a benchmark for more complex classifiers such as support vector artificial neural networks (SVMs). Despite its simplicity, KNN can outperform the most powerful classifiers and is used in a variety of applications such as economic forecasting, data compression, and genetics.
   K-Nearest Neighbors. This algorithm consists of selecting a value of K. At the time of analysis, the K data closest to the value to be predicted will be the solution.
   Here the important thing is to select a value of K according to the data to have a greater precision in the prediction.

3. **The Support Vector Algorithm**

   Support Vector Machine is a discriminant classifier formally defined by a separating hyperplane. In other words, given the labeled training data the algorithm generates an optimal hyperplane that classifies the new examples in two dimensional spaces, this hyperplane is a line that divides a plane into two parts where each class is on each side. Support Vector Machine Support vectors are based on the concept of decision planes that define decision boundaries.

4. **Decision Tree Classification**

   Decision Tree Classification is a type of supervised learning algorithm that is mainly used in classification problems, although it works for both categorical and continuous input and output variables.
   In this technique, we divide the data into two or more homogeneous sets based on the most significant differentiator in the input variables. The decision tree identifies the most significant variable and its value that provides the best homogeneous population sets. All input variables and all possible split points are evaluated and the one with the best result is chosen.
   Tree-based learning algorithms are considered one of the best and most widely used supervised learning methods. Tree-based methods power predictive models with high accuracy, stability, and ease of interpretation. Unlike linear models, they map nonlinear relationships quite well.

5. **Random Forest Classification**

   Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also performs dimensional reduction methods, deals with missing values, outliers, and other essential data exploration steps. It is a type of ensemble learning method, where a group of weak models are combined to form a powerful model.
   Random Forest Classification In Random Forest, several decision tree algorithms are executed instead of a single one. To classify a new object based on attributes, each decision tree gives a classification and finally the decision with the most "votes" is the prediction of the algorithm.

3.1.2 Regression

In this category of supervised learning, instead of classifying the data into different categories, the labels are usually a numerical value and the algorithm consists of trying to approximate a number according to the characteristics of the input data.
It is often used for cases such as weather forecasts, detection of the margin of error when translating a text, etc.

Regression Models in Machine Learning

1. **linear regression**

   linear regression is a method for predicting the dependent variable (y) based on the values of the independent variables (x). It can be used for cases where we want to predict some continuous quantity, for example, predicting the traffic in a retail store, predicting the dwell time of a user or the number of page views in a blog, etc.

2. **Polynomial Regression**

   This algorithm is very similar to the previous one, the difference is that the data here is not linear, so polynomials of degree n must be implemented to obtain the model.
   This algorithm is a trial and error where different degrees of polynomials must be tried to obtain the one that best suits the data. Increasing the degree of the polynomial will not always improve the model, sometimes it makes the algorithm worse, so this is an experimentation process to obtain the most suitable one and reduce the errors between the model and the data.

3. **Support Vector Regression**

   This algorithm is based on finding the curve or hyperplane that models the trend of the training data and, based on it, predicting any data in the future. This curve is always accompanied by a range (maximum margin), both on the positive and negative sides, which has the same behavior or shape of the curve.
   All the data that are outside the range are considered errors, so it is necessary to calculate the distance between it and the ranges. This distance is called epsilon and affects the final equation of the model.

4. **Decision Tree Regression**

   Tree-based learning algorithms power predictive models with high accuracy, stability and ease of interpretation.
   This algorithm is very good at handling tabular data with numerical features or categorical features with fewer than hundreds of categories. Unlike linear models, decision trees can capture the nonlinear interaction between features and target.

5. **Random Forest Regression**
   Random Forest Regression is a supervised classification algorithm. As its name suggests, this algorithm creates the forest with multiple trees.
   In general, the more trees there are in the forest, the more robust the forest will be. Similarly, in the random forest classifier, the higher the number of trees in the forest, the higher the accuracy.
   Once each decision tree has been calculated, the result of each of them is averaged and with this the prediction of the problem is obtained [20].

1.3.2: Unsupervised learning

Unsupervised learning is one that works with unlabeled data, that is, there is no prior knowledge of the type of data. This type of learning tries to find a way to recognize similarities between the data in order to group them according to their characteristics.
The input data has the following structure: Given some input data (X1. . . XN) where X is a vector of D dimensions or characteristics and N is the number of examples.
Due to how this type of learning works, unsupervised learning is often used primarily for data compression or grouping.

3.2.1 Clustering

Clustering is a data mining technique which groups unlabeled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information. Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic [21].

Types

There are different types of grouping that can be used:

- **Exclusive (partition):**
  In this grouping method, data is grouped in such a way that a data can only belong to one cluster or group. Example: K- Means.

- **Agglomerative :**
  In this clustering technique, each piece of data is a cluster. Iterative joins between the two closest clusters reduce the number of clusters. Example: hierarchical grouping.

- **Overlapping :**
  In this technique, fuzzy sets are used to group data. Each point can belong to two or more groups with different degrees of affiliation. Here the data will be associated with an appropriate membership value. Example: Fuzzy C-Means.

- **Probabilistic :**
  This technique uses the probability distribution to create the clusters.

Clustering Models in Machine Learning

- Hierarchical clustering
- K-means clustering
- Mean Shift clustering DBSCAN
- Clustering using Gaussian mixture models
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis

3.2.2 Association

Association rules allow you to establish associations between data objects within large databases. This unsupervised technique tries to discover interesting relationships between variables in large databases. For example, people who buy a new house are more likely to buy new furniture [22].

Association Algorithms

- Prior Algorithm
- Eclat algorithm
- AIS algorithm
- SETM algorithm

1.3.3: Reinforcement learning

Reinforcement learning is a type of learning that has long been used in statistical, neuroscience, and computer science work, but has recently gained a lot of interest in the machine learning environment. This type of learning is a way of programming agents through rewards and punishments without the need to specify from the beginning how to accomplish the task to be learned.
In the reinforcement learning model, the agent in each iteration receives an input "I", an indicator of the current state of the environment "S", with which the agent decides on an action "A" in order to generate the output. This action changes the state of the environment and the value of this state is communicated to the agent through a signal that indicates whether the change is good or bad (Reward or punishment) "R". Over time, the agent would choose the actions that would increase said signal and thus, little by little, it would learn through trial and error.
Because of this, reinforcement learning is often used to teach machines to play games such as chess or robots to behave in a specific way [2].

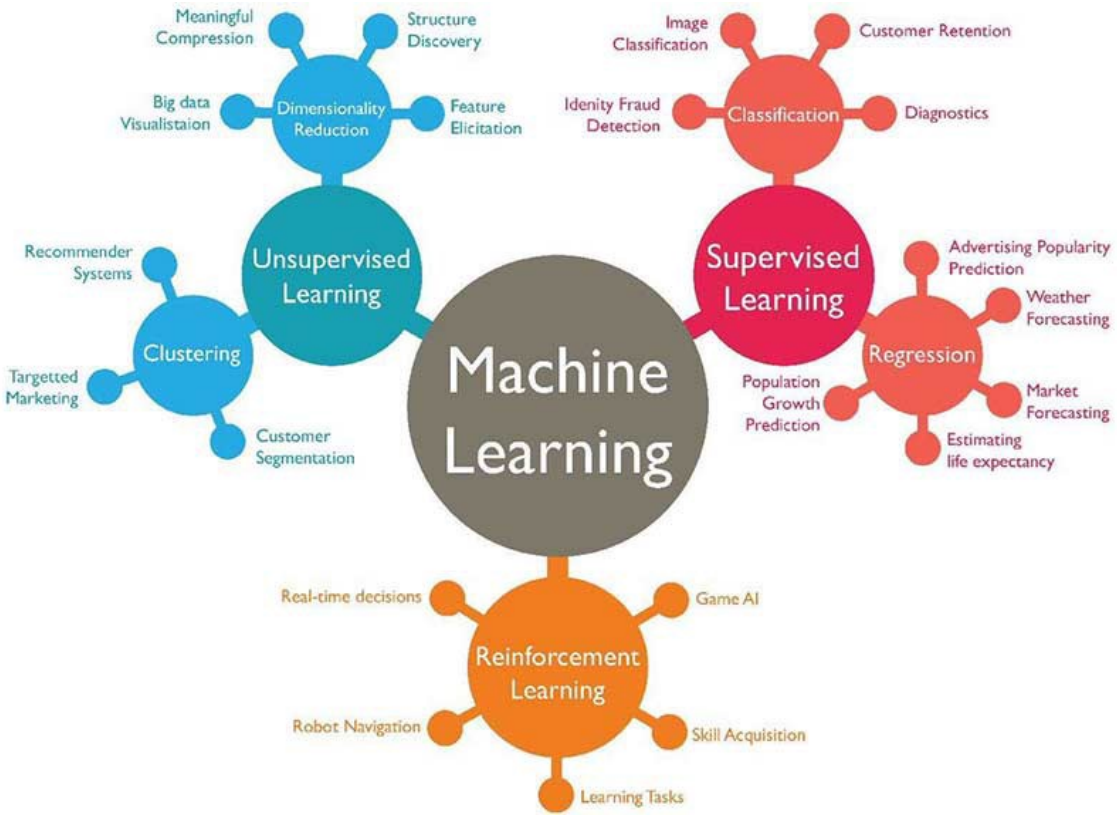

Figure 1.2: Type Of Machine Learning [2].

1.4: Criteria for selecting a machine learning type

Understanding some of the different considerations when choosing a good model is critical to ensuring a successful project.
Let's start with the model's performance and revisit some of the other considerations to keep in mind when selecting a model to solve a problem.

Performance of the model

The quality of the model's results is a fundamental factor to take into account when choosing a model. You want to prioritize algorithms that maximize that performance.
Depending on the problem, different metrics could be useful to analyze the results of the model. For example, some of the most popular metrics include accuracy, precision, recall, and f1-score.
Keep in mind that not every metric works in every situation. For example, accuracy is not appropriate when working with imbalanced datasets. Selecting a good metric (or set of metrics) to evaluate your model's performance is a crucial task before we are ready to start the model selection process.

Explainability of results

In many situations, explaining the results of a model is paramount. Unfortunately, many algorithms work like black boxes, and the results are hard to explain regardless of how good they are.
The lack of explainability may be a deal-breaker in those situations.
Linear Regression and Decision Trees are good candidates when explainability is an issue. Neural networks, not so much. Understanding how easy it is to interpret the result of each model is important before picking a good candidate.
Interestingly, explainability and complexity are usually on both ends of the spectrum, so let's tackle complexity next

Model's complexity

A complex model can find more interesting patterns in the data, but at the same time, it will be harder to maintain and explain.
A couple of loose generalizations to keep in mind:

- More complexity can lead to better performance but also larger costs
- Complexity is inversely proportional to explainability. The more complex the model is, the harder it will be to explain its results.

Putting explainability aside, the cost of building and maintaining a model is a crucial factor for a successful project. A complex setup will have an increasing impact during the entire lifecycle of a model.

The size of the dataset

The amount of training data available is one of the main factors you should consider when choosing a model.
A Neural Network is really good at processing and synthesizing tons of data. A KNN (K-Nearest Neighbors) model is much better with fewer examples.
Going beyond the amount of available data, a related consideration is how much of it you truly need to achieve good results. Sometimes you can build a great solution with 100 training examples; sometimes, you need 100,000.
Use this information about your problem and the amount of data to choose a model that's capable of processing it.

The dimensionality of the data

It's useful to look at dimensionality in two different ways: the vertical size of a dataset represents the amount of data we have. The horizontal size represents the number of features.
We already discussed how the vertical dimension influences the selection of a good model. It turns out that the horizontal dimension is also important to consider: More features will often lead your model to come up with better solutions. More features also increase the complexity of your model.
The Curse of Dimensionality is a great introduction to understand how dimensionality affects the complexity of a model.
As you might imagine, not every model scales the same with high-dimensional datasets. We may also need to introduce specific dimensionality reduction algorithms when high-dimensional datasets become a problem. PCA is one of the most popular algorithms for this.

Training time and cost

How long it takes, and how much it costs to train a model? Would you choose a 98% -accurate model that costs $100,000 to train or a 97% -accurate model that costs $10,000. Of course, the answer to this question depends on your individual circumstances.
Models that need to incorporate new knowledge in near real-time can't afford long training cycles. For example, a recommendation system that needs to be constantly updated with every user's action benefits from an inexpensive training cycle.
Balancing time, costs, and performance is crucial when designing a scalable solution.

Inference time

How long does it take to run a model and make a prediction? Imagine a self-driving system: it needs to make decisions in real-time, so any model that takes too long to run can't be considered.
For example, most of the processing needed to develop predictions using KNN happens during inference time. This makes it expensive to run. However, a Decision Tree will be lighter during inference time and will require more time during training [23].

1. **Categorize the problem:**
   The next step is to categorize the problem.

   - **Categorize by the input:** If it is a labeled data, it's a supervised learning problem. If it's unlabeled data with the purpose of finding structure, it's an unsupervised learning problem. If the solution implies to optimize an objective function by interacting with an environment, it's a reinforcement learning problem..

- **Categorize by output:** If the output of the model is a number, it's a regression problem. If the output of the model is a class, it's a classification problem. If the output of the model is a set of input groups, it's a clustering problem.

2. **Understand Your Data :**
   Data itself is not the end game, but rather the raw material in the whole analysis process. Successful companies not only capture and have access to data, but they're also able to derive insights that drive better decisions, which result in better customer service, competitive differentiation, and higher revenue growth. The process of understanding the data plays a key role in the process of choosing the right algorithm for the right problem.
   Some algorithms can work with smaller sample sets while others require tons and tons of samples. Certain algorithms work with categorical data while others like to work with numerical input.

   - **Analyze the Data:** The components of data processing include pre-processing, profiling, cleansing, it often also involves pulling together data from different internal systems and external sources.
   - **Process the data:** The components of data processing include pre-processing, profiling, cleansing, it often also involves pulling together data from different internal systems and external sources.
   - **Transform the data:** The traditional idea of transforming data from a raw state to a state suitable for modeling is where feature engineering fits in. Transform data and feature engineering may, in fact, be synonyms. And here is a definition of the latter concept. Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. By Jason Brownlee.

3. **Find the available algorithms :**
   After categorizing the problem and understand the data, the next milestone is identifying the algorithms that are applicable and practical to implement in a reasonable time. Some of the elements affecting the choice of a model are:

   - The accuracy of the model.
   - The interpretability of the model.
   - The complexity of the model.
   - The scalability of the model.
   - How long does it take to build, train, and test the model?
   - How long does it take to make predictions using the model?
   - Does the model meet the business goal?

4. **Implement machine learning algorithms :**
   Set up a machine learning pipeline that compares the performance of each algorithm on the dataset using a set of carefully selected evaluation criteria. Another approach is to use the same algorithm on different subgroups of datasets. The best solution for this is to do it once or have a service running that does this in intervals when new data is added.

5. **Optimize hyper-parameters :**
   There are three options for optimizing hyper-parameters, grid search, random search, and Bayesian optimization [24].

1.5: Conclusion

In this chapter, we talked about the concept of machine learning, its importance and applications, the difference between it and the old traditional programming, to reach a census of its most important types, and finally the criteria used to choose the ideal type of machine learning to solve the problem studied. As we know, the right choice means the success of the model.

Chapter 2

Literature review

2.1: Introduction

Since almost the beginning of humankind, men and women have done many things to influence the gender of their offspring for many reasons. Some families would like to choose their baby's gender for family balancing purposes, especially when a family want to add a boy to a long line of girls and vice versa, or selecting to have one child of each gender. For other families, there may be a deep desire for a certain gender for their next newborn in order to avoid to be prone to a specific gender-linked genetic disorder known in their family history such as: Huntington's disease, sickle cell disease, thalassemia, or hemophilia, which means that a child of a given gender would be subjected to a lifetime of medical treatment, pain, and possibly premature death [6]. In addition, in traditional Chinese culture, male heirs are more important and preferred for not only economic reasons but for their ability to continue the family lineage [25]. So, such desire has different reasons and it is neither new, nor is it liable to fade in the foreseeable future.

Actually, and after a mother has a positive pregnancy test result, one of the first and the most exciting questions that comes to her mind is the gender of her future baby. Although there are a few scientific signs that it could offer clues to the fetus gender, there is not a strong scientific answer to this subject before the mother pregnancy. That is why one can find in the literature considerable number of attempts, assumptions and old thoughts about it.

Nevertheless, it is clear that traditional and popular assumptions cannot necessarily reach the same degree of accuracy as scientific researches that seek to determine the gender of the fetus at very advanced stages, but they are often used in conjunction.

In other words, many parents rely on traditional assumptions although they may not guarantee the gender of the fetus by 100%, and by virtue of the fact that relying on it does not have any consequences, but can rather add a kind of pleasure and hope to the subject.

Practically, parents can find out the gender of their kid through an anatomy ultrasound scan at around 18 to 22 weeks of pregnancy, or sometimes with a blood test around week 10 or week 12. However, even professionals may occasionally make the wrong decision about the gender prediction in this period. In addition, the gender of the next newborn will be unclear and it will be too late to discover it in the weeks leading up to birth.

In other words, an ultrasound performed at 13 weeks — normally as part of a specific pregnancy exam — can provide a strong indication of the fetus gender. Furthermore, non-invasive prenatal testing (NIPT), a screening that detects fetal DNA from the baby placenta floating in his mother's blood, can be performed as early as week 10 of pregnancy. However, when the preference for choosing the gender of fetus precedes the conception date, the topic becomes even more beneficial and more exciting, since if some parents are empowered to choose the gender of their newborn, they will be able to spare their children some pain and suffering as cited before [26, 25].

We will present in this chapter a list of some of the most known and interesting gender prediction tests and approaches used by parents all across the world.

2.2: Gender prediction existing works

In recent years, gender prediction problem has occupied an important place in academia and industry as it is become a common issue in machine learning. In fact, the advancement in this research field can contribute to many potential applications and research questions such as: security and surveillance field, E-commerce and demographic research. Practically, a variety of techniques have been presented for this purpose and achieved certain successes. We classify here, some well-known works into two main categories:

2.2.1: General gender prediction

As biological gender is one of the main aspects of presenting individual human, gender prediction is a common topic in machine learning. Hence, much work has been carried out on the classification of gender according to different criteria. We cite here some of those based on machine learning.

1. **Finger print based approaches**

   It is known that fingerprints are one of the common biometric and forensic tools used in personal identification. Therefore, much work has been done on personal and gender identification based on fingerprints.

In this approaches class, we had found [27], where authors cited that some studies in machine learning investigated a correlation between fingerprint and gender. In these studies, by analyzing the fingerprint many important information such as age and gender of a person can be discovered. Many other statistical studies have been made in different geographical areas to identify the relationship between fingerprint and gender. However, in their personal paper authors illustrates gender classification based on fingerprints through various machine learning techniques like naïve Bayes method, Decision Tree and Support Vector Machine algorithms, KNN, PCA, Wilcoxon-Mann-Whitney Test, Friedman Test. In fact, they introduced in their study the concept of the epidermal ridge, minutiae, ridge areas, ridge density etc. And compare above-stated machine learning techniques, their limitations and strengths based on experimental results for gender classification based on fingerprints. As result, this study can be useful for legislative cases and for researchers to devise new machine learning techniques with improved results.

A second study was done in [28] and based on the idea that the process of classifying image can be performing with deep learning where the process like the working of the brain in thinking and trying to reproduce part of its functions by using units associated with relationship, like a neuron. In this research, authors had classified gender based on fingerprint using method Convolutional Neural Network, and then they made three models to determined gender, with a total of 49270 image data that included test data and training data by classifying two categories, male and female. Of the three models, authors had taken the highest accuracy to use in making this application. As results is that interested ones by this research can get Model2 that may be used as a model CNN with the accuracy level of 99.9667%.

2. **Iris based approaches:**

Gender classification based on iris data is a relatively recent concept. in this subsection we try to cite some works that had used artificial intelligence techniques in this context. In [29], a robust iris gender-identification method based on a deep convolutional neural network was introduced. The proposed architecture segments the iris from a background image using the graph-cut segmentation technique. Their proposed model contains 16 subsequent layers; three are convolutional layers for feature extraction with different convolution window sizes, followed by three fully connected layers for classification. The original dataset of this study consists of 3,000 images, 1,500 images for men and 1,500 images for women. They had adopted an augmentation technique to increase the number of dataset images to 9,000 images for the training phase, 3,000 images for the testing phase and 3,000 images for the verification phase that led to a significant improvement in testing accuracy, and the proposed architecture achieved 98.88%. According to their conclusion, the proposed architecture outperformed the other related works in terms of testing accuracy.

In another work [30], authors concluded that learning gender-iris representations through the use of deep neural networks may increase the performance obtained on these tasks. To this end, they had proposed the application of deep-learning methods to separate the gender-from-iris images even when the amount of learning data is limited, using an unsupervised stage with Restricted Boltzmann Machine (RBM) and a supervised stage using a Convolutional Neural Network (CNN). Practically, their trained model had reached 83.00% without data augmentation and 84.66% with data augmentation being one of the top performance on relation with other state of the art methods.

3. **Face image based approaches:**

Over the last years, face analysis methods using machine learning have received immense attention due to their diverse applications in various fields such as security, surveillance and human–computer interaction. In addition, human face image can convey information such as age, gender, identity, emotion and race . . . etc. Therefore, many works have been published that are based on face recognition.

In this context, we cite the research published in [31] where a new algorithm for automatic live Gender Recognition (GR) using Support Vector Machine (SVM) is proposed for classification. The implementation of result work tested on FEI, Live Images and SCIEN database for GR. The detection rate has reached up to 98% in FEI dataset, 94% in Live/Own dataset and 91 SCIEN dataset respectively. Their proposed state of art methodology is compared with the previous techniques and achieved better results which will help in the development of real-time identification systems. In addition, research in [32] is a very good survey in which a comprehensive review focusing on methods in both controlled and uncontrolled conditions is presented. In this work authors illustrated both merits and demerits of each method previously proposed, starting from seminal works on face image analysis and ending with the latest ideas exploiting deep learning frameworks. Finally, they show a comparison of the performance of the previous methods on standard datasets and also present some promising future directions on the topic.

4. **Name based approaches:**
   There has been an increased interest in gender prediction methods based on first names that make use of a variety of open data sources. For instance, in [33], authors focused on simplified Chinese characters and presented a novel approach incorporating phonetic information to enhance Chinese word embedding trained on BERT model. They had compared the proposed method with several previous methods, namely Naive Bayes, GBDT, and Random forest with word embedding via fastText as features. The results show the superiority of their model as they achieved 93.45 test accuracy using their method.
   In addition, they have released two large-scale gender-labeled datasets (one with over one million first names and the other with over six million full names) used as a part of this study for the community.

   Another work is presented in [32], where authors proposed a new dataset for gender prediction based on Vietnamese names. This dataset comprises over 26,000 full names annotated with genders. such dataset is available on their website for research purposes. In addition, they described six machine learning algorithms (Support Vector Machine, Multinomial Naive Bayes, Bernoulli Naive Bayes, Decision Tree, Random Forrest and Logistic Regression) and a deep learning model (LSTM) with fastText word embedding for gender prediction on Vietnamese names. They created a dataset and investigated the impact of each name component on detecting gender. As a result, the best F1-score that we have achieved is up to 96% on LSTM model and researchers generated a web API based on their trained model.

2.2.2: Fetus gender prediction

In the modern era, many systems and application based on artificial intelligence techniques and algorithms are widely used in the domain of fetus gender determination.

2.2.1 Scientific based techniques:

In [34], a common Phonocardiogram (PCG) signal processing techniques was applied on the gender-tagged Shiraz University Fetal Heart Sounds Database in order to study the applicability of previously proposed features in classifying fetal gender using both Machine Learning and Deep Learning models. Since PCG data is generally of contaminated nature with the noise of various types, authors experimented with both static and adaptive noise reduction techniques such as Low-pass filtering, Denoising Autoencoders, and Source Separators. In addition, they have applied a wide range of previously proposed classifiers to the used dataset and they proposed a novel ensemble method of Fetal Gender Identification (FGI). This method substantially outperformed the baseline and reached up to 91% accuracy in classifying fetal gender of unseen subjects.

In another work presented in [35], authors aimed to develop an efficient automated ultrasound-based fetal gender classification model that can facilitate efficient screening and reduce misclassification methods. They have developed a novel feature engineering model termed PFP-LHCINCA that minimizes calculated feature parameter-derived k-nearest neighbor-based misclassification rates. Their model was trained and tested on a sizeable expert-labeled dataset comprising 339 male's and 332 female's fetal ultrasound images. Standard model performance metrics were compared using five classifiers—k-nearest neighbor (kNN), decision tree, naïve Bayes, linear discriminant, and support vector machine (SVM)—with the hyper-parameters tuned using a Bayesian optimizer. The PFP-LHCINCA model achieved a gender classification accuracy of 88% with all five classifiers and the best accuracy rates (>98%) with kNN and SVM classifiers. As conclusions, ultrasound-based fetal gender classification is feasible and accurate using the presented PFP-LHCINCA model.

2.2.2 Popular and traditional techniques

Although there are some scientific ways —as cited before— for predicting the gender of a baby, such as ultrasound scans, many people prefer to employ folk and traditional ways. Such ways range from using the Chinese baby gender chart to examining the mother's-to-be physical and dietary changes after pregnancy. Below we present some of the most popular ways in China for predicting a baby's gender.

1. **Chinese gender prediction method:**

   The "Chinese Gender Calendar" (CGC) is an astrological technique that is traditionally used in China as a gender selection tool, which means that this ancient technique determines the periods when a woman is more likely to conceive a baby of a desired gender. According to this method, the gender of a descendant can often be determined by the Chinese lunar age of the mother at conception and the Chinese lunar month of conception. Despite its popularity, it has been examined by many studies such as [35, 36]. However, most of them, it resulted in an average relationship and there is no predictable influence between the month of pregnancy of the mother and her age at the time of conception with the gender of the newborn.

2. **Mayan gender prediction method:**

   The Mayan gender predictor is another chart so similar to the Chinese one that supposedly able to predict the gender of an unborn child. Such chart is based on two features that are the month of conception from January to December and the mother's age at conception. However, the Mayan gender predictor uses a different methodology to determine its results as described in the following figure Fig.2.1.

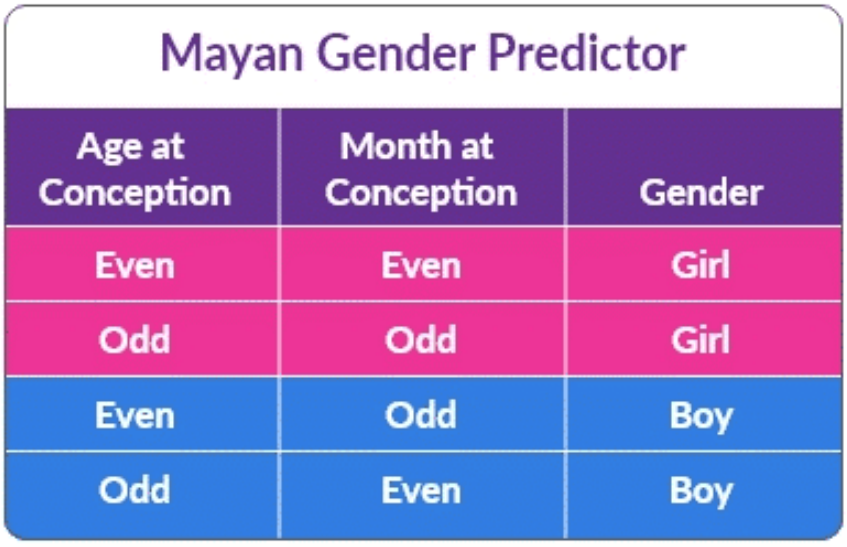

Figure 2.1: Mayan gender predictor chart

In other words, unlike the Chinese gender chart, which needs to find the intersection of the mother's age at the time of conception and the month of conception, the Mayan gender chart only needs to look if the two features are both even or both odd. Therefore, the newborn will be a baby girl, otherwise he will be a boy.

2.3: Summary

The art of attempting to discover the gender determining techniques has been explored for several decades. The ideas expand from folk tales to documented scientific studies. As cited before, avoiding X-linked disease is one of the main motives for discovering the gender of the newborn for a lot of women. Therefore, they try to benefit from technologic advances, traditional and popular methods during childbearing years since some scientific techniques remain very expensive and inaccessible to the general public.

Moreover, we found that certain research that questioned the validity of the Chinese gender table such as [36] lacked adequate data to make such decision, as well as their sharpness in criticism, such as the authors' statement in [37], that accurate prediction of fetal gender based on the Chinese birth calendar was no better than a coin toss, which gives the impression of a non-scientific based decision, and that it is preferable to infer empirical results without such comments. At same time, their study was limited by inclusion of pregnancies resulting from artificial reproductive technologies. However, the Chinese gender table presents the highest likelihood of a normal newborn's gender.

Furthermore, such studies could have been based on an old version of the Chinese table and they missed the fact that it has updated versions across the web periodically.

In conclusion, the adoption of the ancient Chinese and Mayan civilizations on the characteristics of the mother's age at conception as well as the month of conception in an attempt to determine the gender of the newborn calls for scientific curiosity to verify such assumptions in light of the availability of artificial intelligence techniques that can contribute to resolving the controversy about this subject.

Chapter 3

Design and implementation

3.1: Introduction

After getting to know our topic, which is the methods of determining gender across generations, and taking an overview of the field of machine learning, and how to choose the right method for the appropriate method to reach the desired goal. In this chapter, we will define the problem at hand, and then we will present the stages and how to collect and organize data, in order to form the model, train it and evaluate it to achieve the desired goal. Finally, we will briefly describe some tools that help us to achieve these stages, such as Jupyter Notebook, Python language and some of its famous libraries.

3.2: General description of the system

In an attempt to solve our problem, which consists to determine the newborn gender before pregnancy, and after an in-depth study of the existing popular methods, we adopted the machine learning based method due to its encouraging results in such research areas. In addition to the features adopted by Chinese and Mayans of the newborn, we have tried to add the blood groups of the parents and/or any other familial or genetic features, but we were not able to find enough data to cover good learning. Hence, we decided to use the mother age and the month of conception and our system will give as a binary result concerns the newborn gender (male or female). Practically, we have used supervised machine learning and exactly the classification algorithms with the aim to enhance our system with much more features.

3.3: Detailed description

After defining and understanding the problem, we decided to follow the supervised learning by virtue of its meeting the needs of the project and choosing the classification algorithms because we have two choices, they are male or female and the type of classification is binary and from here we go to collect the necessary data to start work.

3.3.1: General architecture

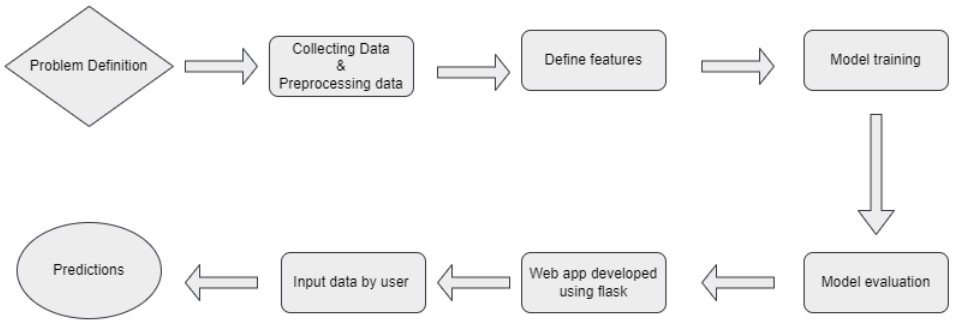

Figure 3.1: System architecture.

3.3.2: Collecting Data

In the course of our search for data, we did not find a set of data on the Internet that would serve us enough. Thereafter, we thought about the idea of a questionnaire, but due to the time limit and fear that most of the answers would not be credible or not enough, resulting in the dispersal of time, as well as the failure of our project. In order to avoid all these problems, we sought state institutions in the hope of satisfactory results.

3.2.1 Dataset source and description

We first found a dataset that includes a primary data source providing birth records data from the Guilford County Register of Deeds from July 1, 2012 to the present time. But it did not give us acceptable results due to the lack of confirmation of the validity of the data. The data was entered randomly, so you find dates of 1453 and 1678, for example, or you find cases without a date. When filtering, we get a small number of data that does not enable us to learn a good one. To face this situation, we did many attempts to find another source of the required data, and finally, we got the required data from reliable and credible institutions (Municipality) and we got nearly 300 thousand cases, which is a good amount for good machine learning.

3.2.2 Dataset features

The initial data contains three features, the date of birth of the mother, the date of birth of the son, and his/her gender. In other words, we had two features of date type and the third is a binary categorical value that is male or female. Of course, such data need to be reformulated in order to prepare the necessary needed features. An example of these data is given as follows:

	A	B	C
1	DOB Mother	DOB Child	Gender
2	22-08-1972	29-11-2010	F
3	20-04-1994	22-05-2011	M
4	14-11-1974	17-06-2003	M
5	18-07-1979	30-06-2018	F
6	26-01-1991	25-05-2017	M
7	23-12-1989	10-08-2008	M
8	02-09-1985	14-11-2002	F
9	18-06-1994	23-04-2021	F
10	23-08-1968	23-08-2008	F
11	18-01-1987	02-06-2005	M
12	27-05-1981	18-07-2018	F
13	29-07-1994	30-10-2013	F
14	25-09-2000	28-12-2021	F
15	27-07-1983	09-11-2009	F

Figure 3.2: Original Dataset features

3.3.3: Preprocessing data

During this stage, we have to perform many operations to prepare the data as follows:

3.3.1 Preparing data (creating new features)

In order to reformulate the original data, the preprocessing stage passed by several stations: First, we have used the son's date of birth to approximately find out the date of the pregnancy from which we need the month of the pregnancy. In fact, we had calculated two pregnancy dates. In the first one we adopted that whole pregnancy period is 9 months and in the second we adopted the 280 days. Thereafter, we have calculated the mother's age at the date of the pregnancy according to the two previous pregnancy dates. In addition, the obtained age had many additional days more than the age' years. Hence, we have rounded the age once by an increase, then by a decrease, and finally by an average in which we rotated it to the nearest, i,e. if the additional days exceeded 180 days. Practically, the last choice was the best after the system testing.

	A	B	C	E	F	G	H	I	K	L	M	N	O	P	Q
2					Minus					Average			Plus		
3	DOB Mother	DOB Child	Gender	DP	MP	MA	gender	Age(D)	M	MP	MA	gender	MP	MA	gender
4	22-08-1972	29-11-2010	F	22-02-2010	02	37	F	13698	6	02	38	F	02	38	F
5	20-04-1994	22-05-2011	M	15-08-2010	08	16	M	5961	3	08	16	M	08	17	M
6	14-11-1974	17-06-2003	M	10-09-2002	09	27	M	10162	9	09	28	M	09	28	M
7	18-07-1979	30-06-2018	F	23-09-2017	09	38	F	13947	2	09	38	F	09	39	F
8	26-01-1991	25-05-2017	M	18-08-2016	08	25	M	9336	6	08	26	M	08	26	M
9	23-12-1989	10-08-2008	M	04-11-2007	11	17	M	6525	10	11	18	M	11	18	M
10	02-09-1985	14-11-2002	F	07-02-2002	02	16	F	6002	5	02	16	F	02	17	F
11	18-06-1994	23-04-2021	F	17-07-2020	07	26	F	9526	0	07	26	F	07	27	F
12	23-08-1968	23-08-2008	F	17-11-2007	11	39	F	14330	2	11	39	F	11	40	F
13	18-01-1987	02-06-2005	M	26-08-2004	08	17	M	6430	7	08	18	M	08	18	M
14	27-05-1981	18-07-2018	F	11-10-2017	10	36	F	13286	4	10	36	F	10	37	F
15	29-07-1994	30-10-2013	F	23-01-2013	01	18	F	6753	5	01	18	F	01	19	F
16	25-09-2000	28-12-2021	F	23-03-2021	03	20	F	7484	5	03	20	F	03	21	F
17	27-07-1983	09-11-2009	F	02-02-2009	02	25	F	9322	6	02	26	F	02	26	F
18	02-05-1973	03-09-1991	M	27-11-1990	11	17	M	6418	6	11	18	M	11	18	M
19	28-11-1975	04-01-2021	F	30-03-2020	03	44	F	16194	4	03	44	F	03	45	F
20	17-02-1977	29-03-2007	M	22-06-2006	06	29	M	10717	4	06	29	M	06	30	M
21	23-03-1984	04-01-2011	M	30-03-2010	03	26	M	9503	0	03	26	M	03	27	M
22	20-01-1986	29-10-2021	M	22-01-2021	01	35	M	12786	0	01	35	M	01	36	M
23	16-12-1984	14-11-2002	M	07-02-2002	02	17	M	6262	1	02	17	M	02	18	M

Figure 3.3: Some of the used excel functions.

Note:

DOB Mother:The day of birth mother

DOB child: The day of birth child

Gender:the gender of the child

DP:pregnancy date

MP:pregnancy month

MA:mother's age

Age(D):age in days

M:extra months for the full year of life

Finally, we give the following data characteristics to end the data description.

	MP	MA	gender
count	300439.000000	300439.000000	300439.000000
mean	6.397315	26.808330	0.516191
std	3.544995	7.847563	0.499739
min	1.000000	16.000000	0.000000
25%	3.000000	19.000000	0.000000
50%	6.000000	25.000000	1.000000
75%	10.000000	33.000000	1.000000
max	12.000000	46.000000	1.000000

Figure 3.4: General description of the data.

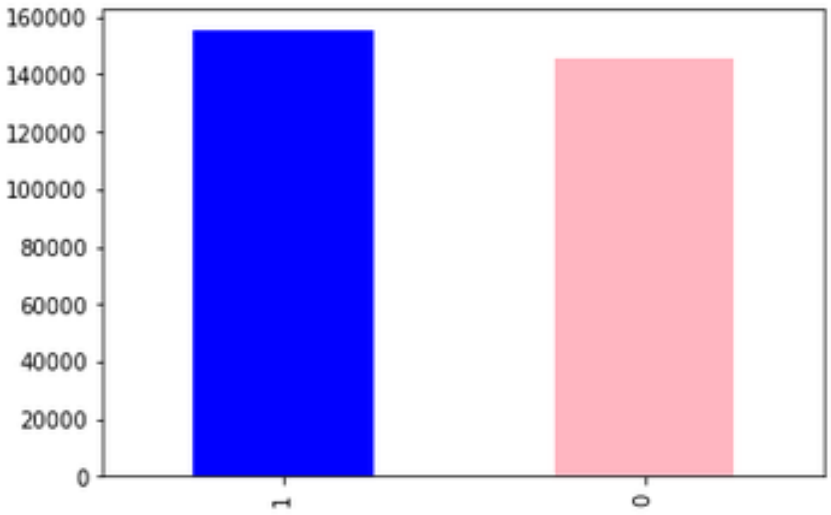

Figure 3.5: Count of elements in each Gender Class.

3.3.2 Filtring data

In the stages of preparing the data, it is expected to find some extreme cases that may impede the good learning such as: the very young/old ages of the mother. Therefore, we think that the most appropriate ages for our study must be between 16 to 46 years. Hence, in our final dataset we have considered only such ages and we found that removing the outliers data increases the accuracy.

3.3.4: Model Building

We briefly give the main steps involved for the model building as follows:

- Setting up features and target as X and Y

```
df.head()
```

	MP	MA	gender
0	2	38	0
1	8	16	1
2	9	28	1
3	9	38	0
4	8	26	1

```
# Split data into X and y
X = df.drop("gender", axis=1)

y = df["gender"]
```

```
X
```

	MP	MA
0	2	38
1	8	16
2	9	28
3	9	38
4	8	26

```
y
```

0	0
1	1
2	1
3	0
4	1

Figure 3.6: The data and split data into X and y.

- Splitting the testing and training sets

```
# Split into train & test set
X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    test_size=0.2)
```

Figure 3.7: Split data into train and test set.

- Build of model for six different classifiers.

  Logistic Regression

  RandomForestClassifier

  KNeighborsClassifier

  Support Vector Machines

  Naïve Bayes

  Decision Tree

3.3.5: Model training

At this stage, we start the process of training model by using our dataset several times according the pre-cited possibilities of mother age and the month of conception features. The result from each cases are stored for comparison reason "model fitting".

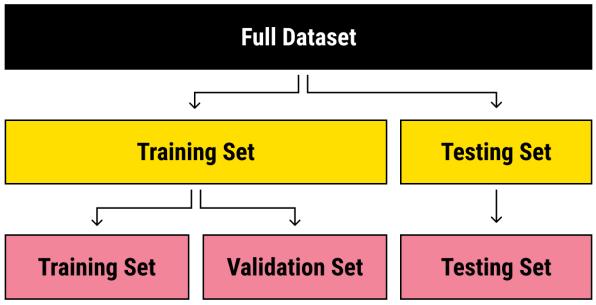

Figure 3.8: Split data in machine learning [3].

3.3.6: Results

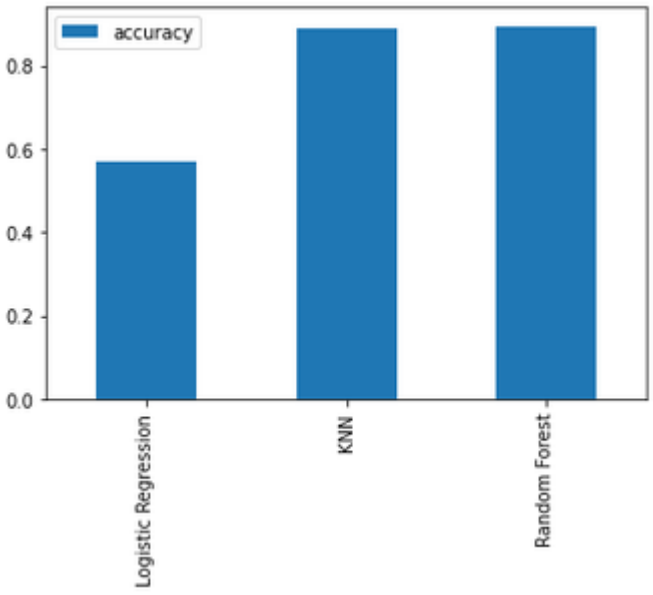

Figure 3.9: Model comparison.

Hyperparameter tuning (by hand)

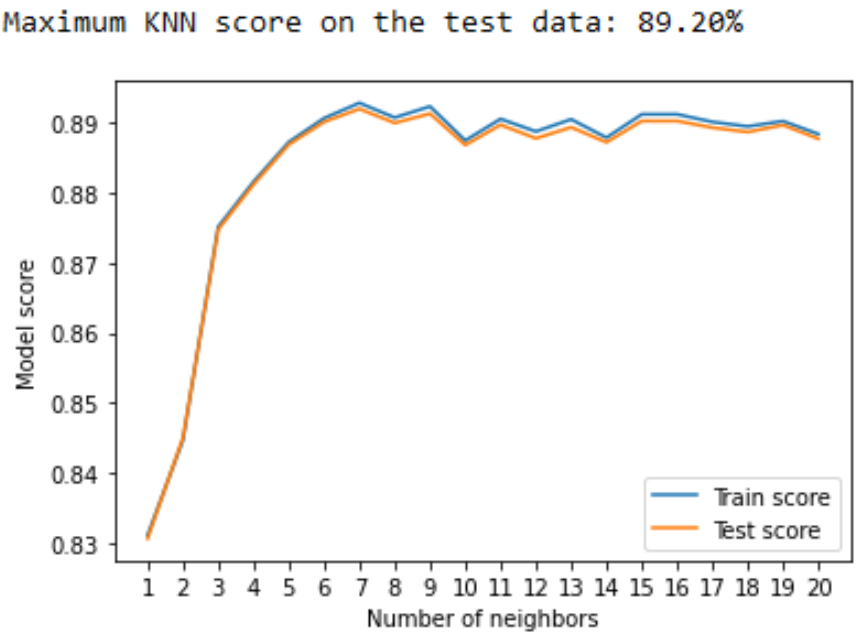

Figure 3.10: Maximum KNN score on the test data.

Hyperparameter tuning with Randomized SearchCV

The de-facto standard for hyperparameter optimization in machine learning has been grid search, which is simply an exhaustive searching through a manually specified sub-set of the hyperparameter space of a learning algorithm. Grid search prevails as the state of the art to hyperparameter optimization due to its simplicity to implement and parallelization, and its reliability in low dimensional spaces. Grid search exhaustively enumerates all combinations of hyperparameters and evaluates each combination. Depending on the available computational resources, the nature of the learning algorithm and size of the problem, each evaluation may take considerable time. Thus, the overall optimization process is time consuming. This leads to an increasing need for efficient methods to optimize hyperparameters. Most of proposed optimization methods tend to reduce the number of evaluations [38].

Area under the Receiver Operating Charavteristic curve(AUC/ROC)

area under curve (AUC)

Receiver Operating Charavteristic curve(ROC)

ROC curves are a comparison of a model's true postive rate (tpr) versus a models false postive rate (fpr).

True postive = model predicts 1 when truth is 1

False postive = model predicts 1 when truth is 0

True negative = model predicts 0 when truth is 0

False negative = model predicts 0 when truth is 1 .

LogisticRegression()

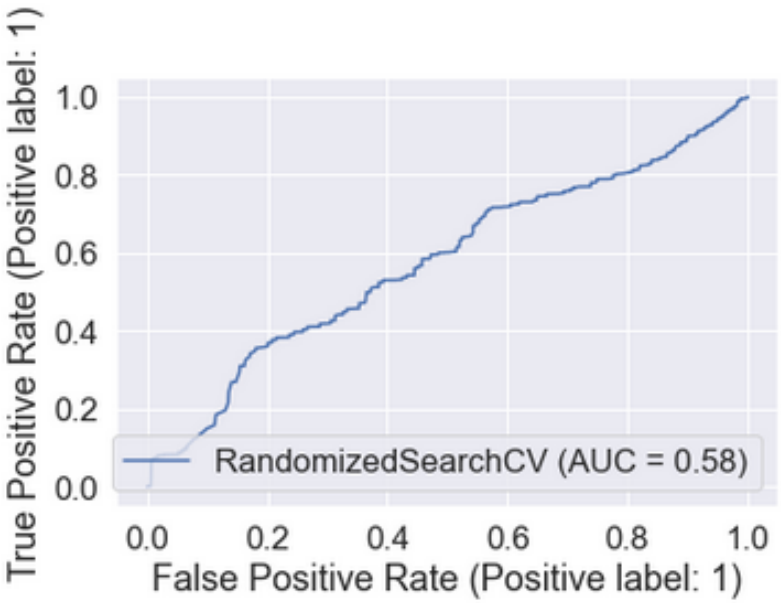

Figure 3.11: AUC Curve of LR.

RandomForestClassifier()

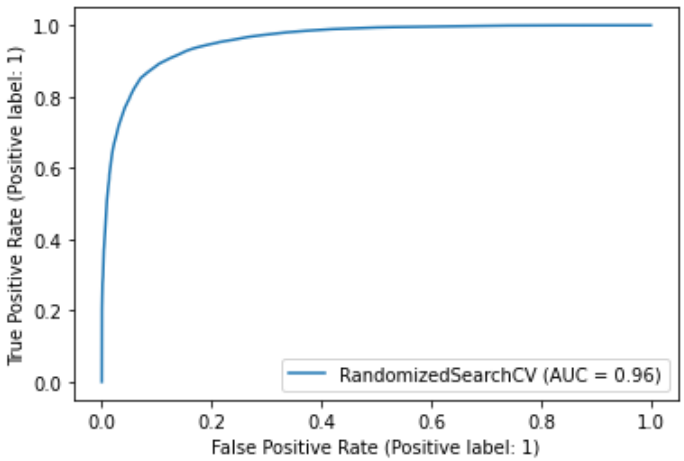

Figure 3.12: AUC Curve of RF.

3.3.7: Model evaluation (with metrics)

Confusion matrix

```
# Confusion matrix
print(confusion_matrix(y_test, y_preds))
```

```
[[26020  3040]
 [ 3336 27692]]
```

Figure 3.13: Confusion matrix of RF.

Classification report

```
print(classification_report(y_test, y_preds))
```

```
              precision    recall  f1-score   support

           0       0.89      0.90      0.89     29060
           1       0.90      0.89      0.90     31028

    accuracy                           0.89     60088
   macro avg       0.89      0.89      0.89     60088
weighted avg       0.89      0.89      0.89     60088
```

Figure 3.14: Classification report.

Accuracy

	Model	Accuracy	AUC
0	Logistic Regression	0.572999	0.57
1	SVM	0.602209	0.60
2	KNN	0.884281	0.88
3	Decision Tree	0.893762	0.89
4	Random Forest	0.893762	0.89
5	Naive Bayes	0.569529	0.57

Figure 3.15: Accuracy results for the algorithms used.

Experimental result In this study, Jupyter Notebook is used to apply the algorithms on the Dataset. Here, the performance is evaluated in terms of accuracy, precision, sensitivity(recall), f1-score, ROC. The factors contributing to these are discussed below:

1. **True Positive** In this scenario, the algorithm predicted true and the actual result is also true.
2. **True Negative** In this scenario, the algorithm predicted false and the actual result is also false.
3. **False Positive** In this scenario, the algorithm predicted true and the actual result is also false.
4. **False Negative** In this scenario, the algorithm predicted false and the actual result is also true.

However, accuracy, sensitivity/recall, precision and f1-score can be calculated using these four outcomes, i.e.

- Accuracy = (True Negative + True Positive)/ (True Negative + False Negative + False Positive + False Positive)
- Sensitivity(Recall) = True Positive/ (True Positive + False Negative)
- Precision = True Positive/ (True Positive + False Positive)
- F1 Score = 2*(Precision * Recall)/ (Precision + Recall)[39].

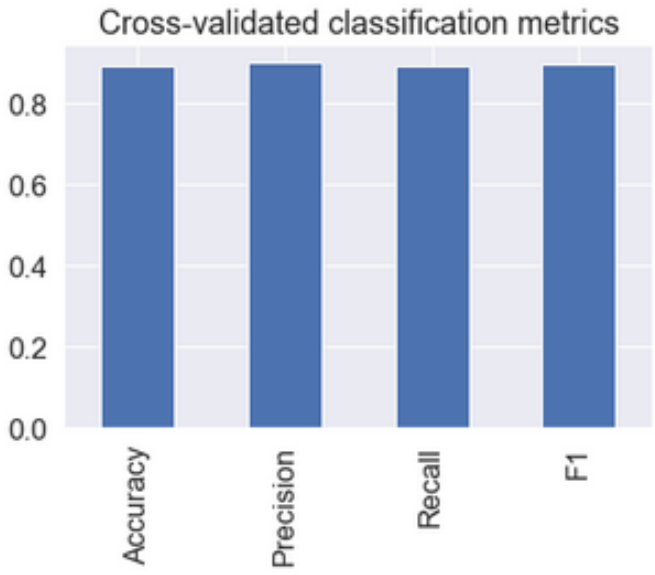

Figure 3.16: Cross-validated classification metrics.

3.3.8: Comparison

Comparison of the accuracy of the algorithms used

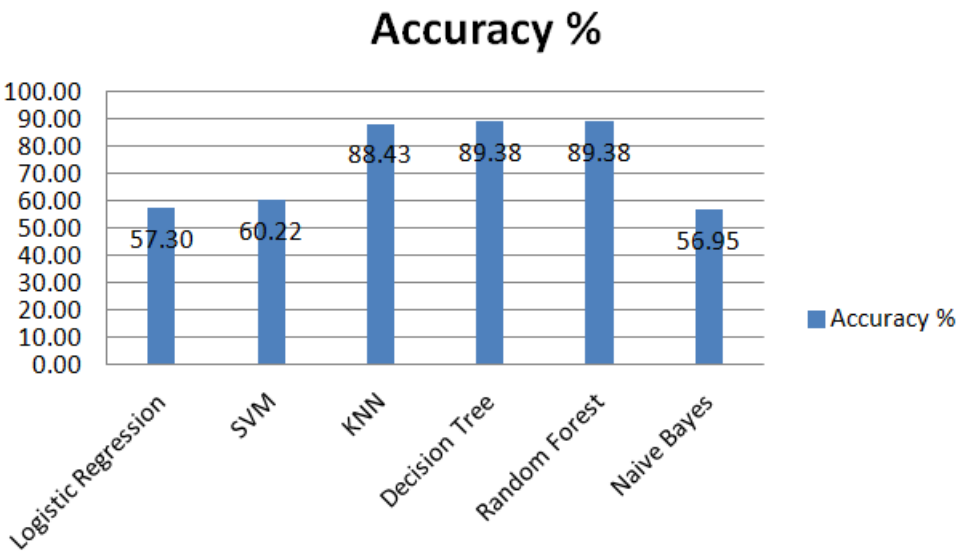

Figure 3.17: Comparison of the accuracy of the algorithms used.

		Minus		Average		Plus	
		9 Months	280 days	9 Months	280 days	9 Months	280 days
LR	All	56.53	56.93	56.74	57.15	56.53	56.93
	18-45	53.95	53.66	55.45	55.88	56.02	56.28
KNN	All	65.87	68.89	80.28	88.69	65.87	68.89
	18-45	66.38	67.99	78.29	87.77	66.93	69.58
Rforest	All	69.78	72	81.84	89.39	69.78	72
	18-45	68.87	71.27	81.61	89.09	69.17	71.9

Table 3.1: Accuracy comparison according to mother'age calculation method

Note: Note from the mother'age account that the best learning cases are average.

the datasets			Average					
			LR		KNN		Rforest	
			9 Months	280 days	9 Months	280 days	9 Months	280 days
1	49000	All	50.51	50.51	50.47	49	50.52	50.4
		18-45	50.08	50.21	50.7	49.9	49.98	50.71
2	160000	All	49.85	51.2	80.07	72.76	83.11	76.35
		18-45	49.15	50.28	79.75	71.54	82.86	75.78
3	253536	All	52.68	52.18	84.5	76.06	86.16	79.17
		18-45	53.92	52.21	84.7	75.49	86.08	79.29
4	300440	All	56.74	57.15	80.28	88.69	81.84	89.39
		18-45	55.45	55.88	78.29	87.77	81.61	89.09

Table 3.2: Accuracy comparison according to data sets volume

S. No.	Attribute	Description	Type
1	MP	Pregnancy Month (in months)	Numerical
2	MA	Mother's Age (in years)	Numerical
3	gender	The gender of the fetus (divided into two categories: 1 for male and 0 for female)	Nominal

Table 3.3: Dataset Description.

Confusion matrix

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	12697	16363	9386	21642
KNN	25438	3622	3174	27854
Random Forest	26020	3040	3336	27692
Naive Bayes	15221	13839	12280	18748
SVM	13119	15941	7877	23151
Decision Tree	26020	3040	3336	27692

Table 3.4: Confusion matrix values for different used algorithms

Analysis of machine learning Algorithm

Algorithm	Accuracy	Precision	Recall	f1-score
Logistic Regression	57.29%	57.01%	70.18%	62.91%
KNN	88.05%	88.86%	87.95%	88.36%
Random Forest	89.44%	90.21%	89.23%	89.71%
Naive Bayes	56.56%	57.48%	60.86%	59.12%
SVM	60.33%	61.32%	60.17%	59.58%
Decision Tree	89.44%	90.21%	89.23%	89.71%

Table 3.5: Evaluation metrics for different used algorithms

3.4: Implementation aspects

To implement the program we needed several means and environments. We mention them as follows:

3.4.1: Python

Python is a platform-independent and object-oriented scripting language, prepared to make any type of program, from Windows applications to network servers or even web pages. It is an interpreted language, which means that the source code does not need to be compiled to be able to run it, which offers advantages such as speed of development and disadvantages such as slower speed.
In recent years the language has become very popular, thanks to several reasons such as:

- The number of libraries it contains, data types and functions built into the language itself, which help to carry out many common tasks without having to program them from scratch.

- The simplicity and speed with which programs are created. A Python program can be 3-5 lines of code less than its Java or C equivalent.
- The number of platforms we can develop on, such as Unix, Windows, OS/2, Mac, Amiga and others.
- Also, Python is free, even for business purposes [40].

3.4.2: Python libraries

Python is a set of libraries that serve different purposes. Among the libraries used to extend our research we cite:

4.2.1 Python visualization libraries

One of the most important stages of the machine learning process is understanding the problem we are going to solve. One way we can improve our understanding of the problem is by understanding the data better. Data visualization helps us better understand the data and the problem.
Similarly, data visualization will also be very useful for understanding the results and analyzing errors. Although there are several Python libraries for data visualization, we will focus on: matplotlib and seaborn for now.

Matplotlib

Matplotlib is the standard and most famous python layout library. You can use matplotlib to create pieces of both the quality needed to publish on paper and in number. With matplotlib you can generate many types of plots: time series, histograms, energy spectra, bar plots, error plots, etc. Matplotlib offers a MATLAB-like environment for creating the best quality graphs and charts for visualizations. In addition, it provides a wide range of functions for developing useful visualizations.

Matplotlib is a standard Python library for creating 2D plots and graphs, it's very low level meaning it requires more commands to generate nice graphs and figures than some more advanced libraries, however the flip side of that is the flexibility, with enough commands, you can do almost any type of plot you want with Matplotlib.

Seaborn

Seaborn is a charting library based on Matplotlib, specializing in statistical data visualization. It features by offering a high-level interface for creating visually attractive and informative statistical graphs.
Seaborn considers visualization an essential aspect when it comes to exploring and understanding data. It integrates well with the Pandas data processing library.

4.2.2 Python Libraries for Numerical Calculation and Data Analysis

Another of the phases of the Machine Learning process that consumes the most time is the preparation of data and the calculation of relevant attributes or characteristics (features). NumPy, and Pandas are the ideal python libraries for data analysis and numerical computation.
Surely we will also face problems that do not require the use of machine learning but only data analysis. Of course, we can also use these libraries in these cases.

Numpy

Numpy is a numerical library that provides universal data structures and high-level mathematical functions.
NumPy provides a universal data structure that enables data analysis and data exchange between different algorithms. The data structures it implements are multidimensional vectors and arrays capable of large amounts of data.
Furthermore, this library provides high-level mathematical functions that operate on these data structures.

Pandas

Pandas is a python library for data manipulation and data analysis Pandas is one of the most useful python libraries for data scientists. The main data structures in pandas are String for one-dimensional data and DataFrame for two-dimensional data.
These are the most widely used data structures in many fields such as finance, statistics, social sciences, and many areas of engineering. Pandas stands out for how easy and flexible it makes data manipulation and data analysis.

4.2.3 Python libraries for Machine Learning

When we looked at the phases of the machine learning process, we saw that building the machine learning model was relatively time consuming. This is so because machine learning libraries already exist. There are several, Scikit-Learn is the most used.

Scikit-learn

scikit-learn is a python library for Machine Learning and Data Analysis. It is based on NumPy, SciPy and Matplotlib. The main advantages of scikit-learn are its ease of use and the large number of machine learning techniques it implements. With scikit-learn we can perform supervised and unsupervised learning. We can use it to solve both classification and regression problems.
It is very easy to use because it has a simple and very consistent interface. The interface is very easy to learn. You realize that the interface is consistent when you can change the machine learning technique by changing just one line of code.
Another point in favor of scikit-learn is that the values of the hyper-parameters have default values suitable for most cases.
Here are some of the machine learning techniques that we can use with scikit-learn:

- linear and polynomial regression
- Logistic regression
- support vector machines
- decision trees
- random forests
- clustering
- instance-based models
- Bayesian classifiers
- dimensionality reduction
- anomaly detection ... etc.

3.4.3: Jupyter Notebook

Jupyter Notebook is a web application for creating documents that contain code, equations, visualizations, and text. You can use Jupyter notebooks to clean data, transform data, perform numerical simulations, statistical modeling, data visualizations, machine learning, and much more.
For practical purposes it is like an interactive python console in a browser that allows you to run python code, display data and graphs, and document what you are doing.
Jupyter is not actually a python library. However, since we are looking at the tools that a data scientist uses the most, the list would not be complete without Jupyter. I use Jupyter constantly to test ideas and build simple prototypes. I do not recommend it when the code is more complex or when we want to create libraries with our work to reuse it in other projects.

3.4.4: Anaconda

Anaconda is a python distribution with the most used libraries by data scientists
Anaconda is a python distribution for Numerical Computing, Data Analysis and Machine Learning. It contains the most used libraries by data scientists. It also makes it very easy to install other libraries that you may need.
With Anaconda it is also possible to create multiple work environments if you are working on multiple projects. This can be useful, for example, if one of the projects requires python 3 and the other python 2. Or if you are working on a project that needs specific libraries or has a specific version.
Unless you have to work with older applications, I recommend that you use the Anaconda distribution with Python 3 [41].

3.4.5: Flask

Flask is a framework that allows you to develop web applications in a simple way, it is specially guided for easy and fast web development with Python. One of its features to highlight is the potential to install extensions or add-ons according to the type of project that you will develop. You only install the functionalities that you are actually going to use [42].

Main used tools and libraries.

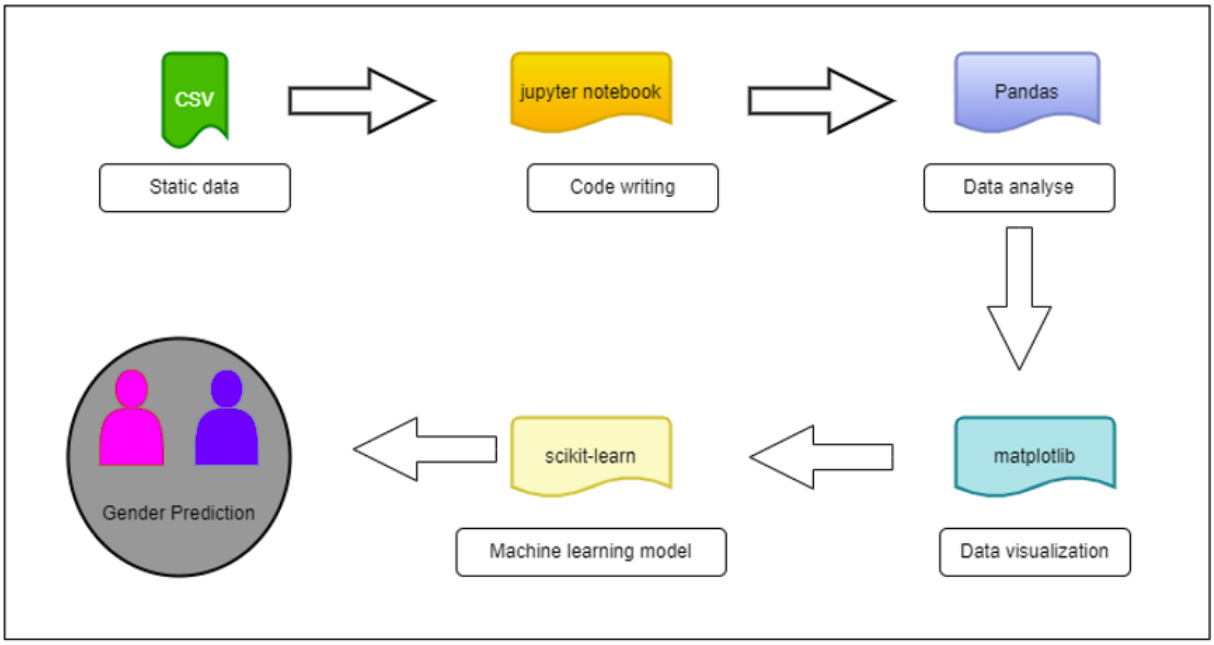

Figure 3.18: Main used tools and libraries.

3.4.6: Some details about implementation

Step 01: Call the necessary Python libraries

```
# Import all the tools we need

# Regular EDA (exploratory data analysis) and plotting libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# we want our plots to appear inside the notebook
%matplotlib inline

# Models from Scikit-Learn
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB


# Model Evaluations
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.model_selection import RandomizedSearchCV, GridSearchCV
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.metrics import precision_score, recall_score, f1_score
from sklearn.metrics import plot_roc_curve

# machine learning model_pipeline

model_pipeline = []
model_pipeline.append(LogisticRegression(solver='liblinear'))
model_pipeline.append(SVC())
model_pipeline.append(KNeighborsClassifier())
model_pipeline.append(DecisionTreeClassifier())
model_pipeline.append(RandomForestClassifier())
model_pipeline.append(GaussianNB())
```

Figure 3.19: Call the necessary Python libraries.

Step 02: Read data file

```
df = pd.read_csv("Average Prepared All Data(280).xlsx - All 280d.csv")
df.shape # (rows, columns)
```

```
(300439, 3)
```

Figure 3.20: Read data file.

Step 03 : Baby's gender prediction system interface

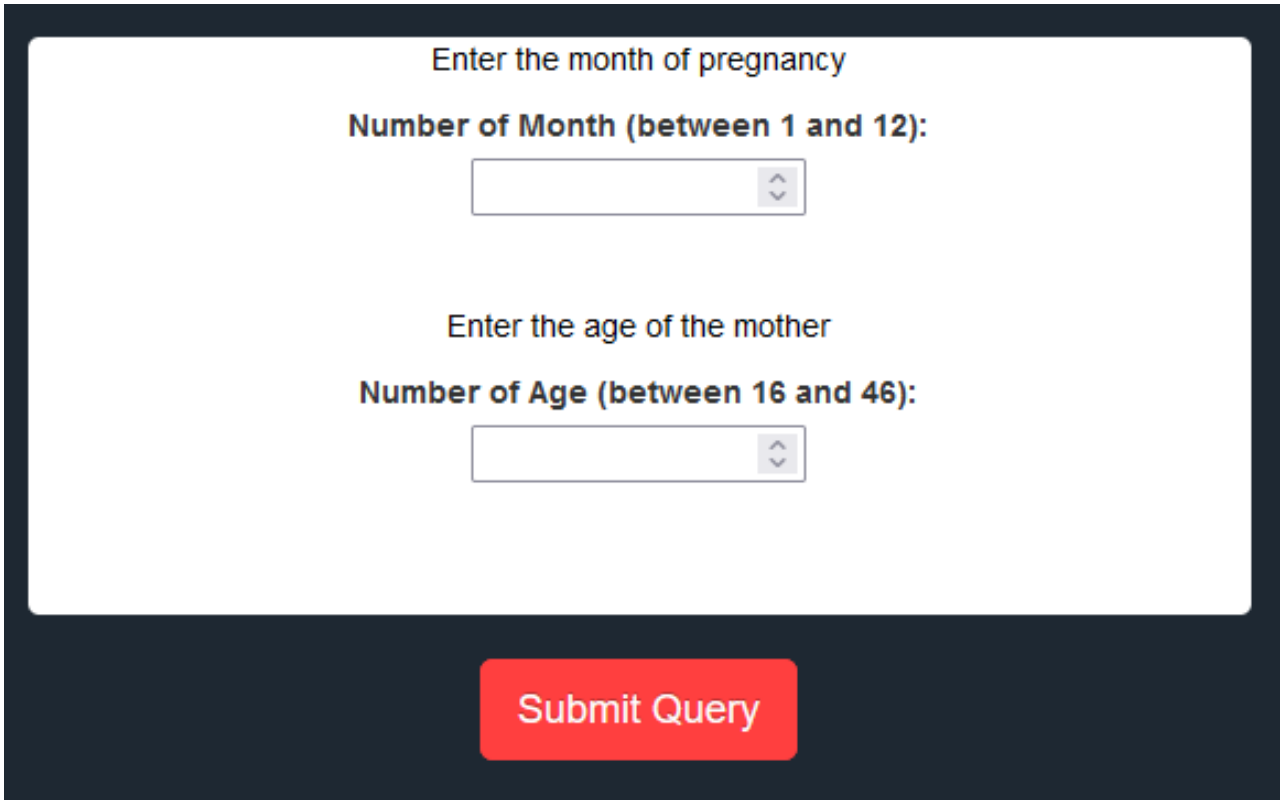

Figure 3.21: Baby'gender prediction system interface.

3.5: Conclusion

Determining the gender of a newborn from some parental data is very difficult even for a doctor, especially early on and that is why many health sectors choose machine learning techniques to determine it. In our experimental study, we took a data set obtained from several reliable institutions and applied preprocessing to drop data with missing values and applied some algorithms such as Logistic Regression, Decision Tree, K-Nearest Neighbor, Support Vector Machines, Random Forest of , which gave Logistic Regression and Naive Bayes and SVM gave an accuracy of 57%, and Decision Tree, Random Forest and K-Nearest Neighbors give an accuracy of 89 %.

In order to facilitate the use of our system for publics, a very simple web widget is created with flask. With such a system, the select the mother's age and the month of pregnancy for which he want the prediction. the data entered by the user can be predicted.

General Conclusion

The ability to predict the gender of a newborn in the future is a topic of great public interest. It has become necessary to develop a system for effectively and accurately predicting the gender of the newborn. The motivation behind the study was to find the most effective ML algorithm for detecting the gender of the newborn. This study compares the accuracy of the logistic regression, Random Forest, KNN, Decision Tree and Naive Bayes algorithms to predict the gender of a newborn using birth registries data from a municipality. The result of this study indicates that the Random Forest and Decision Tree algorithm is the most effective algorithm with an accuracy score of 89.39% for predicting the gender of the newborn.

Future Work

Although the obtained results are very encouraging for scientific research to consider such features, we think that the gender prediction need more and more work. In this context, our work leads to a simple conclusion that it is not strange for a researcher to consider the used features in predicting the gender of newborn since we got a good results, and it is not a 100 percent that there are not other possible factors in this area.
Practically, it is possible to adopt our dataset as a starting point for new projects, even for such research axis or for others such as: early prediction of genetic diseases and so on. At the same time, we aim to extend our study and build a deep learning based model for training larger datasets with more other features such as blood goups and some familial genetic information..

Bibliography

[1] R. ANE, "The Edge Hill University Archive." https://blog.archiveshub.jisc.ac.uk/. Accessed: 2022-01-20. ix, 7

[2] P. Ayush, "Introduction to Machine Learning for Beginners." https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08. Accessed: 2022-03-13. ix, 13, 14

[3] S. Iryna, "Machine Learning and Training Data: What You Need to Know howpublished =https://labelyourdata.com/articles/machine-learning-and-training-data, note = Accessed: 2022-02-23 ." ix, 33

[4] C. Douglas, "Folk tradition and folk medicine in scotland: The writings of david rorie," *BMJ*, vol. 308, no. 6937, p. 1177, 1994. 1

[5] H. Caly, H. Rabiei, P. Coste-Mazeau, S. Hantz, S. Alain, J.-L. Eyraud, T. Chianea, C. Caly, D. Makowski, N. Hadjikhani, *et al.*, "Machine learning analysis of pregnancy data enables early identification of a subpopulation of newborns with asd," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021. 1

[6] A. Abdolrashidi, M. Minaei, E. Azimi, and S. Minaee, "Age and gender prediction from face images using attentional convolutional network," *arXiv preprint arXiv:2010.03791*, 2020. 1, 19

[7] A. L. Chukwumah, "Gender prediction from the primary fingerprint pattern-a study among medical students in ambrose alli university, ekpoma," *Biomedical Journal of Scientific & Technical Research*, vol. 29, no. 1, pp. 22100–22104, 2020. 1

[8] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Speaker gender recognition based on deep neural networks and resnet50," *Wireless Communications and Mobile Computing*, vol. 2022, 2022. 1

[9] Chinese gender predictor, "Baby gender chart 2022." https://www.yourchineseastrology.com/calendar/baby-gender-predictor.htm. Accessed: 2022-06-08. 1

[10] Picpurify, "Online api for face, gender and age detection (demo version)." https://www.picpurify.com/demo-face-gender-age.html. Accessed: 2022-06-08. 1

[11] Voice Gender Recognition, "API/SDK." https://presentid.com/Demo/GenderRecognition. Accessed: 2022-06-08. 1

[12] T. Lisa, "An introduction to machine learning." https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning. Accessed: 2022-06-12. 3

[13] fragmento tomado de The Wall Street Journal, "Aprendizaje automático." https://www.sas.com/es_es/insights/analytics/machine-learning.html. Accessed: 2022-04-08. 3

[14] G. Ligdi, "Definición de Machine Learning." https://aprendeia.com/definicion-de-machine-learning/. Accessed: 2022-03-12. 3

[15] G. Kabanda, *On the Theoretical Foundations of Computer Science. An Introductory Essay*. GRIN Verlag, 2019. 4

[16] R. B.Priya, T. M, and M. A, "Ml-the child of mi," *International Journal of Engineering Technology Science and Research*, vol. 5, pp. 235–237, 2018. 4

[17] S. I. Hosseini, "A quick review on the impact of machine learning and natural language processing on business," in *International Conference of Students, Postgraduates, Young Scientists*, pp. 267–271, Publishing House UMC UPI, 2018. 6

[18] Daffodil Software, "9 Applications of Machine Learning from Day-to-Day Life." https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0. Accessed: 2022-03-18. 7

[19] G. Gibran, "Introducción al machine learning en Python." https://naps.com.mx/blog/introduccion-al-machine-learning-en-python/. Accessed: 2022-03-20. 7

[20] G. Ligdi, "Clasificación de Machine Learning." https://aprendeia.com/clasificacion-de-machine-learning/. Accessed: 2022-02-20. 12

[21] IBM Cloud Education, "Unsupervised Learning." https://www.ibm.com/cloud/learn/unsupervised-learning. Accessed: 2022-02-13. 12

[22] G. Ligdi, "Reglas de Asociación." https://aprendeia.com/reglas-de-asociacion/. Accessed: 2022-03-13. 13

[23] V. Santiago, "Considerations when choosing a machine learning model." https://towardsdatascience.com/considerations-when-choosing-a-machine-learning-model-aa31f52c27f3. Accessed: 2022-0213. 16

[24] A. A. Zaid, "Do you know how to choose the right machine learning algorithm among 7 different types?." https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60. Accessed: 2022-02-14. 18

[25] H.-H. Chen, I.-A. Wang, S.-Y. Fang, Y.-J. Chou, and C.-Y. Chen, "Gender differences in the risk of depressive disorders following the loss of a young child: a nationwide population-based longitudinal study," *BMC psychiatry*, vol. 21, no. 1, pp. 1–12, 2021. 19, 20

[26] R. E. Weiss, *Guarantee the Sex of Your Baby: Choose a Girl Or Boy Using Today's 99.9% Accurate Sex Selection Techniques*. Ulysses Press, 2006. 20

[27] J. S. Yadav, A. K. Gupta, and A. Saxena, "A review on gender identification using machine learning based on fingerprints," *Journal of Information and Optimization Sciences*, vol. 40, no. 5, pp. 1121–1129, 2019. 21

[28] A. I. Gustisyaf and A. Sinaga, "Implementation of convolutional neural network to classification gender based on fingerprint.," *International Journal of Modern Education & Computer Science*, vol. 13, no. 4, 2021. 21

[29] N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and H. N. E. T. Mohamed, "Deep iris: deep learning for gender classification through iris patterns," *Acta Informatica Medica*, vol. 27, no. 2, p. 96, 2019. 21

[30] J. Tapia and C. Aravena, "Gender classification from nir iris images using deep learning," in *Deep learning for biometrics*, pp. 219–239, Springer, 2017. 22

[31] S. Kumar, S. Singh, and J. Kumar, "Gender classification using machine learning with multi-feature method," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0648–0653, IEEE, 2019. 22

[32] M. H. Siddiqi, K. Khan, R. U. Khan, and A. Alsirhani, "Face image analysis using machine learning: A survey on recent trends and applications," *Electronics*, vol. 11, no. 8, p. 1210, 2022. 22, 23

[33] J. Jia and Q. Zhao, "Gender prediction based on chinese name," in *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 676–683, Springer, 2019. 23

[34] R. Khanmohammadi, M. S. Mirshafiee, M. M. Ghassemi, and T. Alhanai, "Fetal gender identification using machine and deep learning algorithms on phonocardiogram signals," *arXiv preprint arXiv:2110.06131*, 2021. 23

[35] A. D. Raj, R. Abichandani, and H. Sethi, "Relationship of lunar phases and sex of the foetus: a retrospective study.," *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, vol. 10, no. 8, pp. 3023–3029, 2021. 24

[36] O. O'Shea, *Validity of the Chinese lunar calendar to predict gender*. Texas Woman's University, 2003. 24, 25

[37] D. Katz and B. Wylie, "568: The chinese birth calendar for prediction of gender-fact or fiction?," *American Journal of Obstetrics & Gynecology*, vol. 201, no. 6, p. S211, 2009. 25

[38] From the journal Open Computer Science, "Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity." `https://www.degruyter.com/document/doi/10.1515/comp-2019-0011/html`. Accessed: 2022-03-23. 34

[39] K. Srivastava and D. K. Choubey, "Heart disease prediction using machine learning and data mining," *International Journal of Recent Technology and Engineering*, vol. 9, no. 1, pp. 212–219, 2020. 37

[40] Miguel Angel Alvarez, "Lenguaje de programación de propósito general, orientado a objetos, que también puede utilizarse para el desarrollo web.." `https://desarrolloweb.com/articulos/1325.php`. Accessed: 2022-02-14. 40

[41] M. H. Jose, "15 Librerías de Python para Machine Learning.." `https://www.iartificial.net/librerias-de-python-para-machine-learning/`. Accessed: 2022-02-20. 43

[42] danestves, "Flask o Django para desarrollo web con Python?." `https://platzi.com/blog/flask-django-desarrollo-web-python/`. Accessed: 2022-02-26. 43