UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

Simulador árboles de decisión

Presentado por Daniel Drefs Fernandes
en Universidad de Burgos — 2 de mayo de 2024
Tutores: Carlos López Nozal
Ismael Ramos Pérez

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

D. nombre tutor, profesor del departamento de nombre departamento, área de nombre área.

Expone:

Que el alumno D. Daniel Drefs Fernandes, con DNI dni, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado título de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 2 de mayo de 2024

Vº. Bº. del Tutor:

D. nombre tutor

Vº. Bº. del co-tutor:

D. nombre co-tutor

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

Índice general

Índice general iii

Índice de figuras v

Índice de tablas vi

1. Introduction 1

2. Project objectives 3

3. Theoretical concepts 5
- 3.1. Decision Trees . . . 5
- 3.2. Entropy . . . 6
- 3.3. Machine Learning . . . 8
- 3.4. Classification . . . 9
- 3.5. Referencias . . . 11
- 3.6. Imágenes . . . 11
- 3.7. Listas de items . . . 12
- 3.8. Tablas . . . 12

4. Techniques and tools 15
- 4.1. Bootstrap . . . 15
- 4.2. D3.js . . . 16
- 4.3. PyCharm . . . 17
- 4.4. Visual Studio Code . . . 17
- 4.5. GitHub . . . 17
- 4.6. Zube . . . 18

4.7. Codacy 19

5. Relevant aspects of the project development **21**

6. Related works **23**

7. Conclusions and future lines of work **25**

Bibliografía **27**

Índice de figuras

3.1. Decision tree example 6
3.2. Descriptive graph of the entropy function 7
3.3. Example of a confusion matrix 10
3.4. Autómata para una expresión vacía 12
4.1. The Bootstrap grid system 15
4.2. Sprint Board example 19
4.3. Codacy quality evolution 20

Índice de tablas

3.1. Herramientas y tecnologías utilizadas en cada parte del proyecto 13

1. Introduction

Decision trees belong to the most popular models in machine learning and data mining, serving as intuitive and interpretable models for decision-making. For solving classification problems, the algorithms IDE3 and C4.5 have significantly contributed to their development and widespread use. IDE3, or Iterative Dichotomiser 3, laid the foundation for decision tree learning by recursively partitioning data based on attribute values, aiming to maximize information gain at each step. C4.5 improved upon IDE3 by handling continuous attributes, missing values, and pruning techniques, enhancing the robustness and accuracy of decision trees. These algorithms may offer efficient classification and regression tasks while offering insights into decision-making processes.

The aim of this project has been to build a simulator in form of a web application for teaching these algorithms in a way that is easy to understand.

2. Project objectives

The primary objective of this project has been to create an interactive and informative web application focused on educating users about decision tree algorithms such as IDE3 and C4.5. For this purpose, it also has to explain the concept of entropy and how it is connected to decision trees. The users are provided with explanations, color-coded and dynamic step-by-step visualizations which make the concepts easier to understand. The web server uses SVG images to display the algorithm's progress and entropy functions.

3. Theoretical concepts

In the following, all the theoretical concepts relevant for the understanding of the project will be explained.

3.1. Decision Trees

A decision tree [6] is a versatile supervised learning algorithm used for classification and regression tasks. Its goal is to predict the value of a variable based on previously processed input. Its hierarchical structure includes a root node and several internal and leaf nodes.

It starts at the root, which represents the feature that best separates the underlying dataset based on a certain criterion. An example would be information gain, which will be explained in a later section. From there branches extend to internal nodes, also called decision nodes. These internal nodes also represent features along with a decision rule that tells us how to further split the data. These features are continuously evaluated until homogenous subsets are created by the leaf nodes. These represent all the possible outcomes of the dataset with each one corresponding to a class label.

Decision tree learning utilizes a divide and conquer approach, iteratively finding the best split points until all or most of the input data is classified.

An example of a decision tree which evaluates whether a person should buy a game or not:

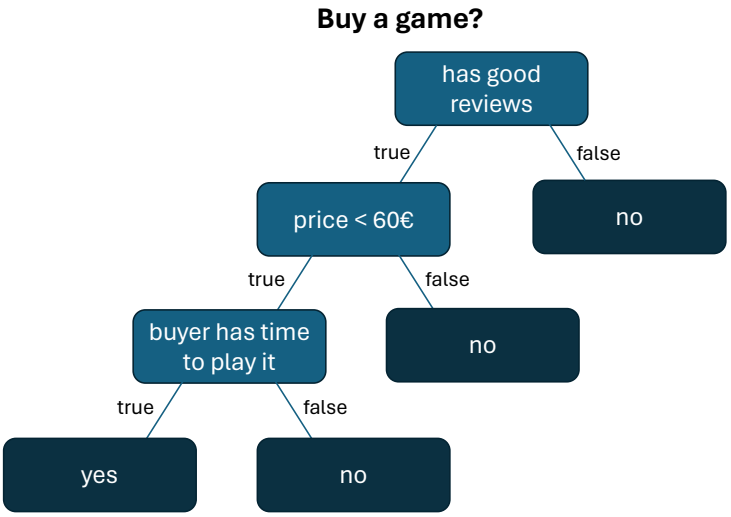

Figura 3.1: Decision tree example

Subsecciones

Además de secciones tenemos subsecciones.

Subsubsecciones

Y subsecciones.

3.2. Entropy

In the context of decision tree algorithms, entropy [5] can be viewed as a measure of impurity in a dataset. It can also be described as a measure of disorder, uncertainty or the expected surprise of a classification, but going forward, the term impurity will be used to avoid confusion. In datasets with binary classes, where variables can only have two possible outcome values, the entropy value lies between 0 and 1, inclusive. The higher the entropy, the more impure the dataset is. In a binary-class dataset, a node that has an equal distribution of, e.g., 5 instances belonging to one class and the other 5 instances belonging to the other class, would have an entropy value of 1. Inversely, a node that has all its instances belong to only one class would have an entropy value of 0, making ´it a pure node. The value is calculated using the following formula:

$$E(X) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

$E(X)$ is the entropy of dataset X, n describes the number of classes in the dataset, and p_i the proportion of instances in class i or, in other words, the probability of an instance belonging to class i.

This following graph [13] that describes the entropy function in relation to the composition of a node shows really well how entropy is used to determine how impure a node is:

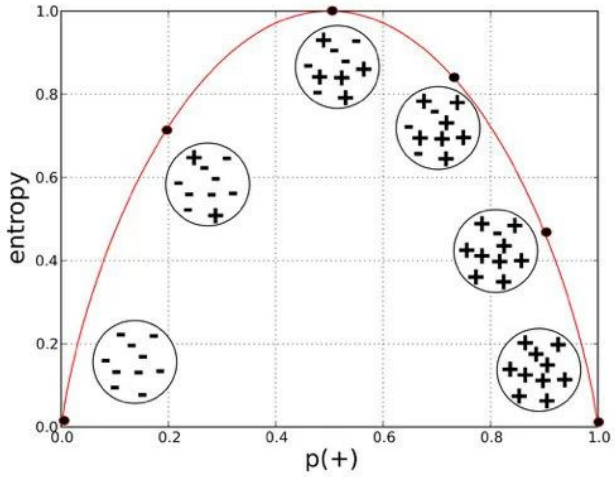

Figura 3.2: Descriptive graph of the entropy function

In a dataset with binary classes, the x-axis describes the amount of instances in a dataset belonging to the positive class while the y-axis measures their entropy values. It records the lowest values at the extremes where there are only or no positive instances recorded in a given dataset while the highest impurity is reached when the numbers of positive and negative instances are equal. However, it must be mentioned that the value of entropy can reach values higher than 1 if the dataset possesses more than 2 class labels.

Conditional Entropy

In the context of decision trees, the conditional entropy [15] $E(Y \mid X)$measures the uncertainty in the target variable Y, which is usually the

class labels, given a certain value of the attribute X that is used for splitting a node. It can be described as the weighted sum of $E(Y \mid X = x)$ for every possible value x of X:

$$E(Y \mid X) = \sum_{x \in X} p(x) E(Y \mid X = x)$$

The formula sums up the conditional entropies for all the possible values x of attribute X. As the weight, it uses $p(x)$, which is the proportion between the number of instances that have the attribute value x and the size of the dataset. Calculating the conditional entropies for the individual values x of attribute X is possible with the following formula:

$$E(Y \mid X = x) = -\sum_{y \in Y} p(y \mid x) log_2(p(y \mid x))$$

$p(y \mid x)$ represents the proportion of instances in the dataset with attribute value x that also have the class value y. It sums over all the possible values of the target variable Y. It can be used synonymous with the formula for $E(X)$ that was previously shown in 3.2

In the end, this conditional entropy calculation evaluates the effectiveness of splitting a node based on the particular attribute X. The lower the conditional entropy, the lower the impurity of the data after splitting the node based on the particular attribute the conditional entropy was calculated for.

3.3. Machine Learning

Machine learning [16] is a field of artificial intelligence which focuses on developing algorithms that are able to learn from given input data and, based on that learned knowledge, generalize to unseen data. Machine learning problems can be categorized into three main tasks: Classification problems, Regression problems, and Clustering problems.

Classification problems

Classification problems [9] in machine learning describe the process of predicting a class or category of a given input. The labeled input data is learned with all its attributes and values to reach the goal of building a model that can also categorize new data, that has not been seen before, into one of the given classes or categories. An example would be classifying emails as spam or not spam. This type of machine learning problem will be discussed in more detail in the section 3.4.

Regression problems

Regression problems [9] in machine learning describe the process of predicting a continuous value based on labeled input data. Instead of predicting a defined class or category, the goal here is to build a model that can predict the quantity of something. An example would be predicting the salary of a person based on their education degree and previous work experience.

Clustering problems

Clustering problems [2] in machine learning do not aim to predict something, like the previous two. Given some unlabeled input data, the goal is to group similar data points together based on certain features to generate an output of clusters inside of which the data of points are more similar to each other than to data points in other clusters.

Decision Trees

With decision trees [6] ultimately following the goal to create a model that predicts the value of a target variable and doing so by learning labeled input data, they are used for both classification and regression tasks.

3.4. Classification

As mentioned before in 3.3, the goal in classification tasks [9] is to assign a class or a label to some input data. Classification aims to create a model that can also generalize to new data. To achieve this, there are multiple algorithms and evaluation metrics that are used to measure a model's performance.

Classification algorithms

K-Nearest Neighbors

K-Nearest Neighbors [8], or short KNN, is an algorithm that uses distance as a measure to determine K of a data point's nearest neighbors. Based on that information, a classification or prediction is made about which of the groups the data point is grouped with. KNN does not go through a training stage, instead it just stores and memorizes the training dataset.

Support Vector Machine

Support Vector Machine [17], or short SVM, is an algorithm that can perform linear as well as non-linear classification. Linear classifiers assume that a data set is linearly separable while non-linear classifiers are not bound to that restrictive belief. With SVM being able to perform both methods, it is able to find a hyperplane that best separates different classes in a data set by maximizing the margin between classes.

Logistic Regression

This algorithm [7] is mostly used for binary classification problems. It uses the logistic function to determine the probability of a given input belonging to a certain class.

Evaluation metrics

Confusion Matrix

A confusion matrix [1] is a table that is used to describe a classification model's performance. It presents an overview of the predictions on the test data set against the actual classes:

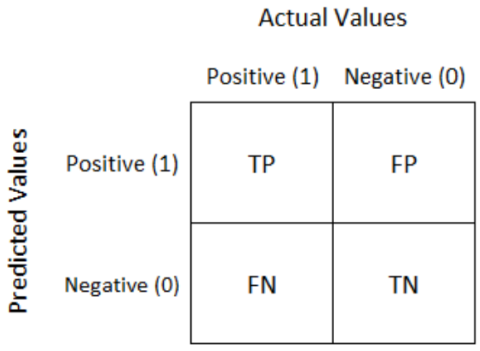

Figura 3.3: Example of a confusion matrix

Where: TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative. It is useful for measuring other metrics like Accuracy, Precision and Recall

Accuracy

Accuracy [1] measures how many of all the classified instances were given the correct label. It uses the following formula:

$$Accuracy = TP + TN/TP + TN + FP + FN$$

This metric is useful for well balanced data sets. If 99 of the data set's instances belonged to one class and only 1 belonged to the other, likeliness is high that the used model would always predict the class with the higher amount of instances, resulting in a 99 % accuracy due to the data set's composure.

Precision

Precision [1] describes the ratio of correctly predicted positive cases to the total predicted positive cases. In short words, it measures the accuracy of the positive class predictions and is calculated with the following formula:

$$Precision = TP/TP + FP$$

Precision is useful for cases where False Positives are a bigger concern than False Negatives

Recall (Sensitivity

Recall [1] describes how many of the actual positive cases our model was able to correctly predict. It measures the ability of the model to find all positive instances and is calculated using the following formula:

$$Recall = TP/TP + FN$$

Recall is useful when False Negatives are of a higher concern than False Positives.

3.5. Referencias

Las referencias se incluyen en el texto usando cite [**?**]. Para citar webs, artículos o libros [10], si se desean citar más de uno en el mismo lugar [3, 10].

3.6. Imágenes

Se pueden incluir imágenes con los comandos standard de LaTeX, pero esta plantilla dispone de comandos propios como por ejemplo el siguiente:

Figura 3.4: Autómata para una expresión vacía

3.7. Listas de items

Existen tres posibilidades:

- primer item.
- segundo item.

1. primer item.
2. segundo item.

Primer item más información sobre el primer item.

Segundo item más información sobre el segundo item.

-

3.8. Tablas

Igualmente se pueden usar los comandos específicos de LaTeXo bien usar alguno de los comandos de la plantilla.

Herramientas	App AngularJS	API REST	BD	Memoria
HTML5	X			
CSS3	X			
BOOTSTRAP	X			
JavaScript	X			
AngularJS	X			
Bower	X			
PHP		X		
Karma + Jasmine	X			
Slim framework		X		
Idiorm		X		
Composer		X		
JSON	X	X		
PhpStorm	X	X		
MySQL			X	
PhpMyAdmin			X	
Git + BitBucket	X	X	X	X
MikTEX				X
TEXMaker				X
Astah				X
Balsamiq Mockups	X			
VersionOne	X	X	X	X

Tabla 3.1: Herramientas y tecnologías utilizadas en cada parte del proyecto

4. Techniques and tools

In this section, all the development tools that have been used to carry out the project are presented. This ranges from frameworks like Bootstrap to version control tools like GitHub. Besides a short introduction, their most notable functions and benefits to the project's development are documented.

4.1. Bootstrap

Bootstrap [14] is a front-end framework which is known for providing useful and easy-to-use HTML and CSS templates like tables, buttons, forms and many others. It also comes with JavaScript components like modal dialogues and dropdown menus.

One of Bootstrap's key features is its grid system which lets users divide their web page's contents into rows and columns:

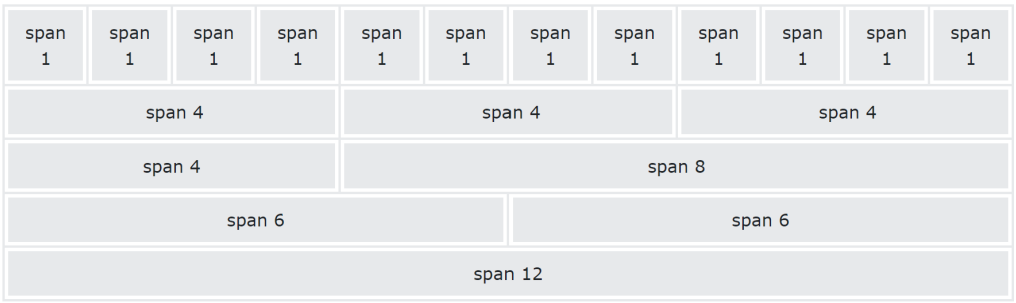

Figura 4.1: The Bootstrap grid system

Each row possesses 12 columns which the user can freely utilize to organize the contents that are to be displayed. Bootstrap also automatically puts the device on which the web page is displayed into one of six different size categories, based on the device's screen width. This, in combination with the grid system, allows the developer to make their website responsive to different screens, ranging from large desktop monitors to smartphones.

In addition to Bootstrap's ability to create responsive websites, the wide variety of CSS classes used for common HTML elements like buttons offer further simplicity to the web design aspect of creating a page. With this, maintaining a visual consistency throughout the project is made easier, too.

In this project, the newest version of Bootstrap at present, Bootstrap 5, is used. The main advantages are its usage of vanilla JavaScript for its components instead of relying on jQuery, new components for better customization and simplified CSS which reduces file size and loading times for the created pages.

4.2. D3.js

"D3-[4] stands for "data-driven documents.ªnd perfectly describes the free, open-source JavaScript library. "Documentsrefers to the Document Object Model (DOM). D3.js lets the user bind data to its elements. The library works like a toolbox that uses a variety of discrete modules which, e.g., allow selection and transition operations. It binds these modules together so all the necessary tools are at hand, ready to be applied.

D3.js does not invent new data presentation formats, instead, it makes use of web standards like SVG to display contents. Incorporating these standards, the library also allows the use of external stylesheets which can be employed to change the graphics' visual representations.

A major feature of D3.js is its ability to dynamically change the displayed contents. Whether that change is triggered by user interactions or a change in underlying data, the library's data join concept allows separate operations for entering, updating and exiting existing DOM elements based on a given set of data. Besides filtering and sorting, it lets you control what happens to your contents in many ways when changes happen and update your website accordingly.

4.3. PyCharm

PyCharm [12] is the Integrated Development Environment (IDE) I used during the first 2 and part of the 3rd sprint. It is an IDE developed by JetBrains which is specifically designed for Python development and is the one I had mainly been using for university assignments. With its support for web development frameworks like Flask and features like code completion for HTML, CSS and JavaScript, it served useful in allowing a quick start into this project's development. PyCharm's built-in live preview for HTML files and its interactive debugging feature also made working on the prototypes much simpler. Its version control integration of systems like Git allowed an easier experience of making local changes remotely available.

However, one of the Issues during the 3rd sprint was to start using Bootstrap to make the website's layout responsive. Due to set-up problems with the IDE, PyCharm was not able to provide auto-completion for the framework. To make use of that and other features, I started using a different program for developing the website during the 3rd sprint.

4.4. Visual Studio Code

Visual Studio Code [11] is an open-source code editor by Microsoft. While it also provides all the benefits mentioned in section 4.3, it comes with a handful of other advantages of which the most significant one would be the vast variety of available extensions. These include Bootstrap IntelliSense which enables CSS class auto-completion, Live Server which launches a local server with a live reload feature, and GitLens which, e.g., allows the user to see inline information about when and in which commit the current line of code was last changed. These extensions allow the user to add features based on what they need. By only including essential coding features upon first installation, the editor take up less space than IDEs like PyCharm. In addition, its lightweight design that is optimized for performance leaves behind a smaller memory footprint.

Ultimately, these features made me switch to Visual Studio Code for the remaining sprints.

4.5. GitHub

GitHub is a platform that is mainly used for version control on software development projects. Not only does it provide a location for users to

store their files, make and track changes, it also makes it possible to share repositories with other developers and therefore presents an easy way to collaborate on projects.

Project management is made easier with GitHub, too. Providing tools for code review, milestone and issue tracking, it gives participants of the project insight on the current progress. Communication through comments on specific issues, commits, pull requests or code reviews is possible, too, allowing for organized feedback.

In this project, GitHub has been used to store all relevant files and update the development in this repository. To be able to set up defined tasks and track progress, these issues have been set up. Furthermore, this GitHub Pages repository has been used to deploy prototypes for different functionalities of the web application. GitHub Pages is a site hosting service that allows the user to host a website directly from a GitHub repository. Doing that made it easier to review sprints with the tutors as the deployed web application could simply be opened on any browser without having to launch it from an IDE or a code editor.

All these functions have been used by the presenter of this work and the tutors to ensure an organized and transparent development process.

4.6. Zube

Zube is a project management platform for software development teams that comes with tools that help to plan and manage projects efficiently. Allowing a direct link to GitHub repositories, issues are synced so that when a change is made to those linked issues on GitHub, that information is transferred to Zube, and vice versa.

Tracking progress is made easy with the inclusion of sprints and allowing the user to add issues to a sprint board which can be used for project management methodologies like Scrum. Charts like Burndown and Burnup show each sprint's progress over the time set for the sprint. The Burndown charts were used to analyze the temporal planning throughout the project in the corresponding section in the Anexos part.

As seen in the image below, the Sprint Board, among other tools, has been used to plan out each sprint thoroughly and document each step in regards to the created issues.

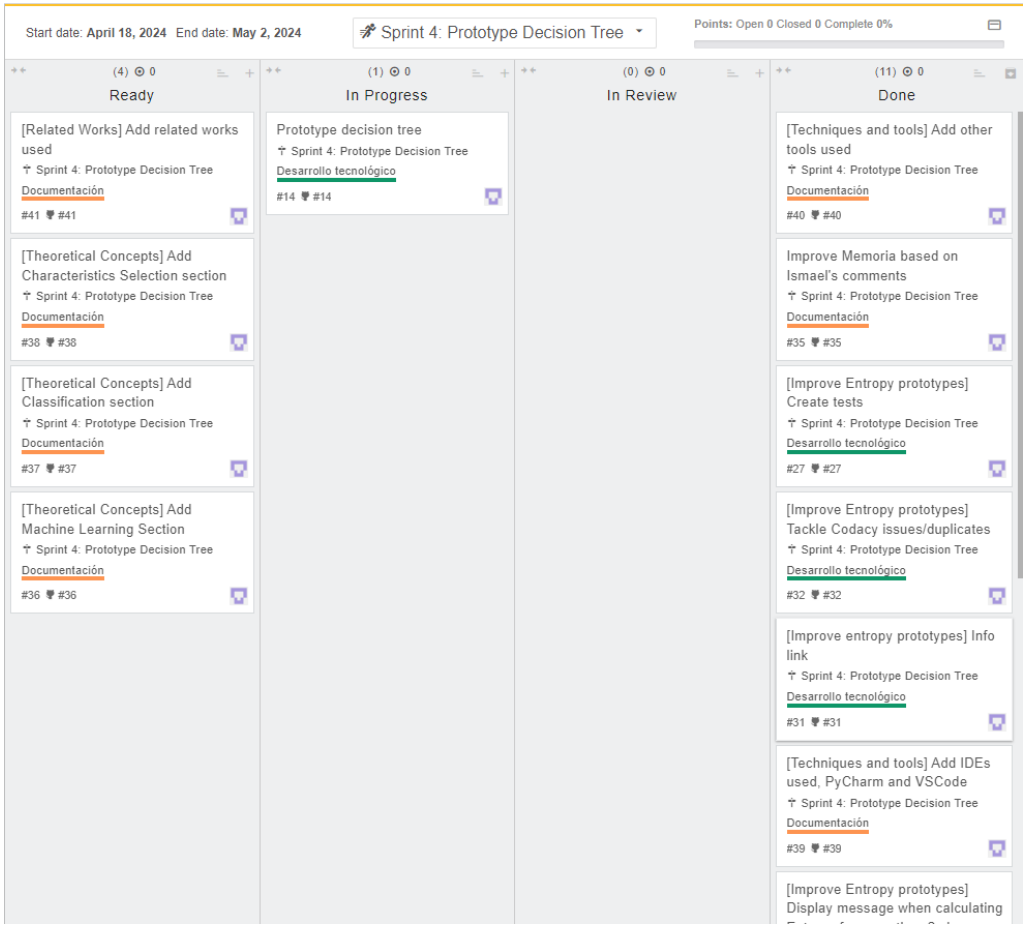

Figura 4.2: Sprint Board example

4.7. Codacy

Codacy is a code analysis tool that helps software developers improve their code quality. Allowing a link to GitHub, it reviews commits in real-time and checks the code's quality based on code security, code duplication, coding style and other general issues. Codacy's coding standards can be customized to fit the developers' own quality standards by making it possible to, e.g., ignore certain issues that are irrelevant to the project, introduce other security guidelines and enforce specific coding styles of their own.

As seen in the image below, which shows the timeline of existing issues within the code in the remote repository, Codacy has been used to continuously identify and tackle all sorts of problems in the code that arose during the development of the web application.

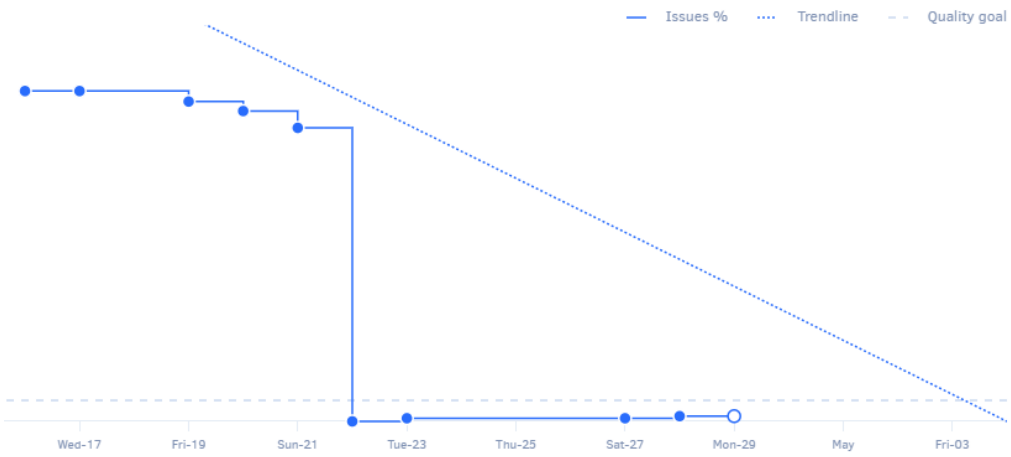

Figura 4.3: Codacy quality evolution

5. Relevant aspects of the project development

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros3, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

6. Related works

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

7. Conclusions and future lines of work

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

[1] Sumeet Kumar Agrawal. Metrics to evaluate your classification model to take the right decisions. https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/, 2024. [Internet; visitado 02-mayo-2024].

[2] Moez Ali. Clustering in machine learning: 5 essential clustering algorithms. https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms, 2022. [Internet; visitado 30-abril-2024].

[3] Zachary J Bortolot and Randolph H Wynne. Estimating forest biomass using small footprint lidar data: An individual tree-based approach that incorporates training data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(6):342–360, 2005.

[4] Mike Bostock and Inc. Observable. What is d3? https://d3js.org/what-is-d3, 2024. [Internet; visitado 15-abril-2024].

[5] Shailey Dash. Decision trees explained — entropy, information gain, gini index, ccp pruning. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c#:~:text=In%20the%20context%20of%20Decision,only%20pass%20or%20only%20fail., 2022. [Internet; visitado 29-marzo-2024].

[6] IBM. What is a decision tree? https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,

internal%20nodes%20and%20leaf%20nodes., -. [Internet; visitado 15-marzo-2024].

[7] IBM. What is logistic regression? https://www.ibm.com/topics/logistic-regression, 2024. [Internet; visitado 02-mayo-2024].

[8] IBM. What is the k-nearest neighbors (knn) algorithm? https://www.ibm.com/topics/knn#:~:text=The%20k-nearest%20neighbors%20(KNN,used%20in%20machine%20learning%20today., 2024. [Internet; visitado 02-mayo-2024].

[9] Zoumana Keita. Classification in machine learning: An introduction. https://www.datacamp.com/blog/classification-machine-learning, 2022. [Internet; visitado 30-abril-2024].

[10] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[11] Microsoft. Code editing. redefined. https://code.visualstudio.com, 2024. [Internet; visitado 21-abril-2024].

[12] JetBrains s.r.o. The python ide for data science and web development. https://www.jetbrains.com/pycharm/, 2024. [Internet; visitado 21-abril-2024].

[13] Sam T. Entropy: How decision trees make decisions. https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8, 2019. [Internet; visitado 29-marzo-2024].

[14] W3Schools. Bootstrap 5 tutorial. https://www.w3schools.com/bootstrap5/index.php, 2024. [Internet; visitado 15-abril-2024].

[15] Wikipedia. Conditional entropy. https://en.wikipedia.org/wiki/Conditional_entropy, 2024. [Internet; visitado 29-marzo-2024].

[16] Wikipedia. Machine learning. https://en.wikipedia.org/wiki/Machine_learning, 2024. [Internet; visitado 30-abril-2024].

[17] Wikipedia. Support vector machine. https://en.wikipedia.org/wiki/Support_vector_machine, 2024. [Internet; visitado 02-mayo-2024].