esprit
Se former autrement
HONORIS UNITED UNIVERSITIES

Data Science Project Report (2022-2023)

Competency-based curriculum platform

Realised By:

Sana Bouhaouala
Yoser Walha
Malek Zommit Chatti
Melek Abid
Omar Nouri
Bilel Kammoun

Class: 4DS5

ECOLE SUPÉRIEURE PRIVÉE D'INGÉNIERIE ET DE TECHNOLOGIES

Abstract

Education plays a crucial role in shaping our future and enhancing our employability. The curriculum we follow during our educational journey serves as a foundation for acquiring knowledge, skills, and competencies necessary to thrive in the job market. However, the rapidly evolving nature of industries and the demands of the 21st-century workforce requires a continuous update of the skills that we are learning. Every year, the universities must find the perfect balance between what they teach their students and the current job market needs, to ensure their future.

This data science project aims to address this challenge by developing a platform that generates the perfect competency-based curriculum tailored to several specialties for universities. Additionally, the platform offers a unique feature where students can upload their resumes for evaluation, comparing them to the current needs of the job market. By providing a comprehensive score and suggesting specific skills for improvement, the platform empowers students to enhance their employability and bridge the gap between their education and industry requirements. Through data-driven insights, this project aims to revolutionize education, preparing students for a successful future in the ever-changing job landscape.

The competency-based curriculum generation component of the project implements data science techniques and algorithms to analyze a wide range of data sources (scraped from the indeed platform), including job market trends, industry demands, and educational resources. By leveraging machine learning and statistical analysis, the platform can identify the essential skills and competencies required for each specialty or field of study. It will also take into consideration the industry demands and future job market projections.

The project also utilizes deep learning techniques, such as Natural Language Processing (NLP) and Optical Character Recognition (OCR), to assess the relevance and proficiency of skills mentioned in a student's resume, comparing them with the prevailing needs of the job market. The evaluation process considers factors such as skill gaps, emerging job roles, and industry-specific requirements to generate a comprehensive score and personalized skill enhancement recommendations.

The proposed platform has the potential to revolutionize university education by aligning curricula with the needs of the job market and providing students with valuable insights into their employability. By offering personalized skill enhancement recommendations, it empowers students to proactively develop the competencies most sought after by employers. That way, the hiring process will be simplified for students.

Keywords: Web Scraping, CRISP-DM, Machine Learning, Deep Learning, Statistical Analysis, OCR, NLP, Data Analysis, Modeling, Classification Models

Table of Contents

List of Tables

List of figures

1. INTRODUCTION

Before online learning got in the game, on-campus learning was the only efficient learning format. The traditional education has been designed for generations now to prepare students for either further education or to enter the job market. It is characterized by a lecture-based approach to teaching: Students typically attend classes in a physical classroom, where they listen to lectures, complete homework assignments, and take exams. This type of education aims to help students memorize information, knowledge or concepts, which can then be applied in a work environment. This is also reflected in the tests or exams, where future graduates need to reproduce what they've learned. Traditional education has been successful in providing students with a basic foundation of knowledge and skills and has been instrumental in advancing knowledge and shaping societies for centuries. However, the focus on rote learning and memorization can lead to a shallow understanding of the material and can limit the creativity and critical thinking skills of students.

Nowadays, our world presents challenges in terms of rapidly advancing technology, globalization, and diverse students in terms of both background and learning needs. Thus, while there is a drive to increase what children learn and know by the time they graduate college, there is also a need to prepare them to compete in a global economy. That's why, we now consider that traditional education may not be sufficient to prepare students for the demands of the 21st century. In fact, the need to transform the education system is being increasingly recognized, with schools across the United States exploring new approaches to education. These approaches are focused on providing more personalized learning experiences, while ensuring that every student has a strong foundation in the skills and competencies required to succeed in today's world. A key approach in this regard is competency-based education (CBE). The fundamental goal of CBE is to allow each student to advance academically, based on their specificability or competency to master a particular skill. The approach allows students to progress at their own pace, with the education delivery being tailored to different learning abilities. So, the basic idea is to design the curriculum based on the student's mastery of a topic, rather than the time taken to master it. The evaluation methods seek to answer the question: How proficient is this student in this skill? It's not about remembering definitions or (abstract) concepts. It's all about showing how you can apply what you know to solve real-world problems. A competency-based curriculum is a curriculum that emphasizes what learners are expected to do rather than mainly focusing on what they are expected to know.

Unfortunately, most of the future graduates have been raised in a traditional educational environment for the time being. Besides, most of their first applications for a specific job will probably be rejected. This is due to their lack of experience and mastery of the flagship skills of the moment. This is why an independent competency-based education is now required. Before applying, each student should look for, and learn the skills they need to maximize their chances of being hired for the job of their dreams.

2. METHODOLOGY USED

We decided to use the CRISP-DM methodology because it provides a structured, step-by-step approach to data mining projects, which helps us to ensure that all necessary steps are taken and that our project stays on track. The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. It has six sequential phases, each with its own objectives, tasks, and deliverables:

1. Business understanding : This phase involves defining the project objectives and goals, and determining the requirements from a business perspective for a successful outcome

2. Data understanding : In this phase, the data sources and data requirements for the project are identified and analyzed, and the initial data is collected and explored.

3. Data preparation: During this phase, the collected data is transformed and cleaned to make it suitable for analysis. This may include handling missing or inconsistent data, dealing with outliers, and normalizing the data.

4. Modeling : This phase involves building and testing data models to identify patterns and relationships in the data. The models are then evaluated and the best one is selected.

5. Evaluation: In this phase, the selected model is evaluated and compared to the project goals and objectives, and its accuracy and effectiveness are measured.

6. Deployment: This is the final phase, where the results of the data mining project are communicated, and the solution is deployed into the production environment. This phase may also involve monitoring the deployed solution and making any necessary improvements over time.

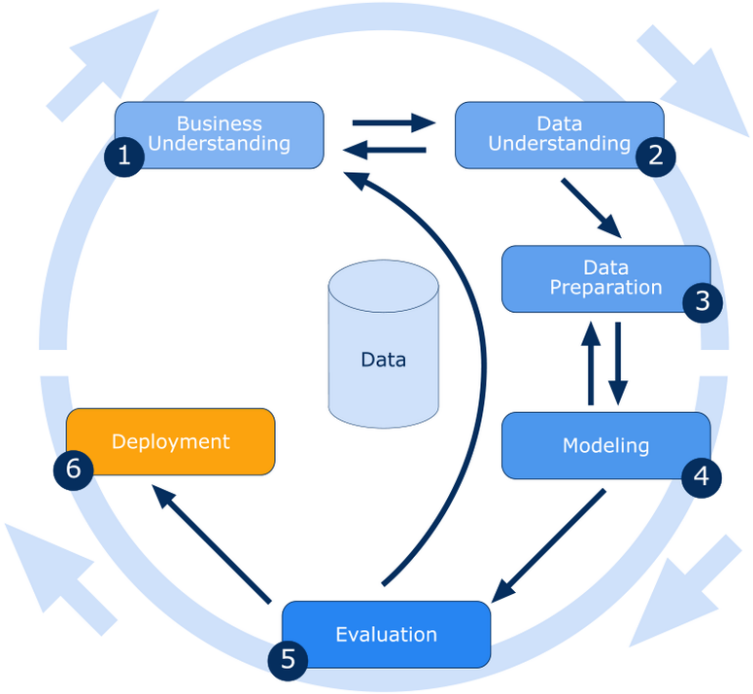

Figure 1: The CRISP-DM Methodology

3. BUSINESS UNDERSTANDING

3.1. Review of the existing competency-based curriculum platforms

In order to help the graduates, find the skills they need to learn, various competency-based curriculum platforms have been developed. Among them:

- Khan Academy: it is a non-profit organization that provides a free, self-paced curriculum in a variety of subjects, including math, science, history, and computer programming. It allows students to progress through the material at their own pace and provides detailed analytics and reporting to track student progress.
- Apex Learning: it is a leading provider of digital curriculum for high schools. It offers a wide range of core, elective and Advanced Placement courses that are designed to be flexible and personalized.
- Edgenuity: it is a digital curriculum provider that offers a wide range of courses for middle and high schools, including core subjects, electives, and Advanced Placement courses. It uses data and analytics to personalize the learning experience for each student and track their progress.
- Knewton: it is an adaptive learning platform that uses data to personalize the learning experience for each student. It offers a wide range of courses for students of all ages, including math, science, and language arts.
- Compass Learning: it provides a personalized learning platform for K-12 students. It offers a wide range of courses, including core subjects, electives, and Advanced Placement courses, and uses data and analytics to personalize the learning experience for each student

These platforms are very helpful for graduates. However, they don't necessarily help them find a complete curriculum that will help them achieve their goals. These platforms will just provide the students different courses on different skills that might be useful to them. Besides they won't be helpful for universities, which are looking forward to giving the best education to their students.

3.2. Solution

A competency-based curriculum platform is an educational system that focuses on the acquisition and demonstration of specific knowledge, skills, and abilities. This type of platform is designed to give students more control over their learning by allowing them to progress through material at their own pace, rather than being tied to a traditional classroom schedule.

Some key features of a competency-based curriculum platform include:

1. Clear learning objectives: The platform should clearly define the knowledge, skills, and abilities that students are expected to acquire.
2. Self-paced learning: Students should be able to progress through the material at their own pace, rather than being tied to a traditional classroom schedule.
3. Personalized learning: The platform should be able to adapt to the individual needs of each student, providing them with the resources and support they need to be successful.

Moreover, a competency-based curriculum projects can have several benefits for businesses, including:

1. Skilled workforce: Businesses need employees with specific knowledge and skills, and competency-based curriculum projects can help to ensure that students are graduating with the skills and abilities that are needed in the workforce.
2. Greater Efficiency: Businesses can benefit from the self-paced nature of competency-based curriculum projects, as employees will be able to progress through training and development programs at their own pace and be ready to work on projects as soon as they finish, improving the overall efficiency of the company.
3. Improved employee retention: Businesses can benefit from improved employee retention by hiring employees who have demonstrated the specific knowledge and skills that are needed for the job, rather than just those who have completed a traditional degree program.
4. Greater flexibility: Competency-based curriculum projects can provide greater flexibility for businesses, as they can be designed to meet the specific needs of the company, rather than being based on a one-size-fits-all approach.

3.3. Target Market

With our competency-based curriculum platform, we are aiming at 2 major targets:

- First, universities: We want to help them provide the students with the "perfect" curriculum to improve the university program and replace the traditional courses. The goal is to help them always be in line and have a modern program to train competent students for the future.

- Then, students/employees: By carrying out this process, the latter are in a strong position to know the skills required for the job or the internship and to be hired. They are also advantaged to choose the complete and perfect profile for their desired mission.
In addition, we facilitate the procedures, we help them save time and energy for precise knowledge and we drive them by placing them in the map they want.

3.4. Functional and non-functional requirements

3.4.1. Non Functional Requirements

Performance requirements refer to the specific outcomes or goals that a system, product, or service is expected to achieve. These requirements may vary depending on the specific context, but some common examples of performance requirements include:

1. Speed: For systems that process large amounts of data or handle many users, performance requirements may include the maximum number of requests that can be handled per second or the time it takes for a request to be processed.

2. Reliability: For systems that need to be available 24/7, performance requirements may include a maximum allowable downtime or a required level of redundancy to ensure that the system can continue to operate even if one or more components fail.
3. Accuracy: For systems that process data, performance requirements may include a minimum level of accuracy for the data being processed, such as the percentage of correct answers in a classification task

GUI (graphical user interface) requirements refer to the specific features and functionality that a GUI should have in order to meet the needs of its users. Some common examples of GUI requirements include:

1. User-friendly: The GUI should be easy to use and understand, with a consistent and intuitive layout, clear labels and instructions, and minimal complexity.
2. Responsiveness: The GUI should respond quickly and smoothly to user input, with minimal delay or lag.
3. Compatibility: The GUI should be compatible with different devices, operating systems, and browser types, and should be able to adapt to different screen sizes and resolutions.
4. Security: The GUI should have built-in security features, such as user authentication and encryption, to protect against unauthorized access and data breaches.

3.4.2. Functional Requirements

Functional requirements refer to the specific features and functionality that a system or product must have in order to meet the needs of its users. Some common functional requirements for a competency-based curriculum platform might include:

1. CV Parsing: our platform should be able to automatically extract relevant information from the candidate's CV, such as education, work experience, skills, and certifications.
2. Skill Matching: Our platform should be able to match the skills and experience listed in the candidate's CV with relevant courses and learning paths.
3. Recommendation engine: Our platform should use algorithms and machine&deep learning models to recommend the most relevant courses and learning paths based on the information in the candidate's CV.
4. Curriculum generator: Our platform should use algorithms, machine learning models and statistical analysis to generate a complete curriculum for each specialty and suggest it for the universities. This curriculum should be accredited by the CTI (Commission des Titres d'Ingénieur).

3.5. Data science requirements

In order to develop our competency-based curriculum platform, we are going to use different techniques such as:

- Web Scraping with Scrapy : Collecting a lot of data is crucial in order to create a platform that universities and students can rely on. Thus, we will scrapy as many Resumes/Job offers and existing curriculums as we can.

- OCR (Optical Character Recognition): This will help us read textual information directly from digital documents and scanned documents (probably CVs) without any human intervention.

- NLP (Natural Language Processing): It is a subfield of Artificial Intelligence (AI) that deals with the interaction between computers and human languages. The goal of NLP is to enable computers to understand, interpret, and generate human languages in a way that is similar to how humans do it. This involves analyzing and understanding the meaning and context of human language, as well as the structure and grammar of the language. This technique will help us extract the keywords to simplify the classification of the competencies needed in each field nowadays.
For example, we can use NLTK (Natural Language Toolkit), which is a popular Python library for NLP, to train and test our NLP models. It is an open-source library that provides a wide range of tools and resources for working with human language data, such as text processing, tokenization, stemming, lemmatization, part-of-speech tagging, syntactic parsing, semantic parsing, and more.
Also, we can use spaCy, which is another popular Python library for NLP that is specifically designed for production use, providing fast and efficient tools for advanced NLP tasks, such as tokenization, part-of-speech tagging, named entity recognition, and dependency parsing.

- Machine & Deep Learning: Surely, we will need machine learning and deep learning algorithms to develop our curriculum recommendation system.

3.6. KPIs

Our project won't be successful if we don't evaluate it throughout the process. That's why we need KPIs. KPI stands for Key Performance Indicator, which is a measurable value that demonstrates how effectively our competency-based curriculum platform is achieving its objectives or goals. KPIs are crucial to evaluate progress, identify areas for improvement, and monitor the success of our models. We have set three important KPIs:

- First, a Skill Evaluator: We will evaluate the skills acquired from a curriculum through a Machine Learning model. The goal is to predict whether these skills correspond to a good profile or not. The decision will be made through the analysis of a database which contains the skills of the engineers and an evaluation variable (the evaluation of the quality of engineer will be taken by his salary).

- Second, a Curriculum Evaluator: We will also evaluate a curriculum through a Machine Learning model which is used to predict whether it is a good or bad course. For that, we will use a database which contains the curriculum of a given school and an output variable which makes the evaluation of this course (the evaluation will be taken by the international rank of the school).

These two KPIs will give us the opportunity to evaluate our curriculum through the skills acquired and the reality of the study plan. As long as the results obtained are not satisfying, we must continue to improve our work. Once we get a well-evaluated curriculum the job will be done.

- Third, a **Feedback analysis:** Once the platform has generated an appropriate curriculum for a field, we need to make sure that the program proposed is successful. That's why, we need to take into consideration the feedbacks of our clients to know if there are changes that need to be made.

4. DATA UNDERSTANDING

4.1. Data Collection

Data collection is the process of capturing both qualitative and quantitative information from various sources. It is one of the most important steps in any research or analysis project, such as a data science project. Without data, we actually can't perform any analysis or make any predictions. Moreover, it is crucial to ensure that the data obtained is accurate and reliable in order to guarantee that the analysis and decision-making processes are founded on valid and credible information. Data may be gathered using a variety of techniques, including surveys, interviews, observations, experiments, and web scraping. We used 3 types of techniques to gather information for this competency-based curriculum platform.

At the beginning of this phase, we mailed surveys to several Esprit alumni. This method was clearly not effective since we gathered only 23 observations from it, and most of them weren't really reliable. However, we weren't very surprised. It is very well known that data collection from forms is often a challenging task as it can be difficult to get a high response rate.

Due to this too few responses issue, we needed to consider alternative solutions for this step. First, we thought that web scraping from professional networking platforms such as LinkedIn could be a useful approach. This involves collecting data from public profiles and resumes of individuals who work in the targeted field. Using this approach, a significant amount of data may be gathered rapidly and easily. However, after a few failed attempts and a threatening email received from the Linkedin firm, we discovered that scraping data from this networking platform is explicitly prohibited by the company's user agreement and can even result in legal consequences. Besides, when researching this topic, we came across a regulation that forbids scraping resumes from any sort of website, without the person's agreement.

Finally, since web scraping resumes on LinkedIn (and any other networking platform) was not a viable option, we decided to collect data on job postings and requirements. The platform "Indeed", for example, is a popular job search engine that can help us harvest data. That's why, we decided to gather data from job postings on this platform using ApiFy. Apify is a web scraping and automation platform that enables developers to extract data from websites, automate workflows, and create custom APIs. It provides a range of tools and services that simplify the process of web scraping.

After collecting the 12 datasets (each specialty has a dataset), we merged them into one single dataset and we added a column, named "Field", in which we specified the specialty (For example: "Data Science" if the observations came from the Data Science dataset that we retrieved).

4.2. Data Storage

4.2.1. Types of databases

- Relational databases: This is the most common type of database, where data is stored in tables with rows and columns. Relational databases use SQL (Structured Query Language) to manipulate and retrieve data.

- NoSQL databases: These databases use a non-relational approach to data storage, which means they don't use tables with fixed columns. Instead, data is stored in various formats, including key-value pairs, documents, and graphs.
- Object-oriented databases: This type of database stores data as objects, which can be manipulated using object-oriented programming languages like Java and Python.
- Hierarchical databases: This type of database stores data in a tree-like structure with parent and child relationships. Each record can have only one parent but multiple children.
- Network databases: Similar to hierarchical databases, network databases store data in a tree-like structure but allow for multiple parent-child relationships.
- Graph databases: These databases store data in nodes and edges, allowing for complex relationships between data.
- Time-series databases: These databases are designed to store time-stamped data, such as sensor data or financial transactions, and make it easy to analyze and retrieve data based on time intervals.
- Spatial databases: These databases are optimized for storing and querying data with a geographic component, such as maps, GPS data, and satellite imagery.

4.2.2. Types of datasets

- Image datasets: These datasets contain images and are commonly used for tasks like object detection, classification, and segmentation.
- Video datasets: Similar to image datasets, video datasets contain frames of video and are used for tasks like action recognition, object tracking, and video classification.
- Text datasets: These datasets contain text data and are used for tasks like sentiment analysis, text classification, and natural language processing (NLP).
- Speech datasets: These datasets contain audio recordings of speech and are used for tasks like speech recognition and speaker identification.
- Reinforcement learning datasets: These datasets are used for training agents in reinforcement learning, which involves maximizing a reward signal in a dynamic environment.

4.2.3. Type of our database

Our dataset is a tabular dataset because data is organized into a table consisting of rows and columns, where each row represents a specific instance or record (an observation), and each column represents a specific attribute or property of that record (a feature). Finally, we can say that we have a Relational Database because it’s stored in tables. Tabular dataset is often stored in spreadsheet software such as Excel or Google Sheets. These datasets can be easily sorted, filtered, and manipulated using spreadsheet functions or external tools, and can be exported to various file formats for use in other applications.

4.3. Data Comprehension

Our database of job offers is very rich with approximately 11900 observations divided into all the IT specialties covered. There are almost 1000 observations of specific job offers for each IT specialty taught in ESPRIT (TWIN, SIM, DS, ERP-BI, INFINI, ArcTic, SAE, SE, IOT, WIN, SLEAM and NIDS). We have tried to grant each option a thousand observations in order to obtain a balanced and rich database at the same time.

By observing the features of the database, we find:

Name of the Feature	Description of the Feature
Position name	Name of the job position offered. It is granted to the requested IT specialty
Description	Rich paragraph of the skills required for this job. From this feature we can extract the courses/skills required for the specialty
Salary	The salary for each job
Rating	Mean rating value for the job offer
Reviews count	Number of comments related to a specific job offer. We can value this feature by taking into consideration the importance of this job and automatically the importance of skills
Location	The country/state of the company that published the job offer
Company	The company that published the job offer
Company Logo	Link to the logo of the company that published the job offer
External Apply Link	Link to apply for the job offer
Id	Unique number to identify the job offer
Job Type	Type of contract of the job offer
Posted At	Number the days since the job offer was published
Scraped At	Date and hour of scraping
Url	Link of the job offer
Field	Field of the job offer (Data science, TWIN ….)

Table 1 : The features of our database

5. DATA PREPARATION

Data preparation is a crucial phase of a data science project that involves transforming and cleaning raw data into a format that can be analyzed and used for modeling. It allows us to explore, clean, combine, and format data for sampling and deploying models. This phase is critical for the success of our data science project, as the quality of the analysis and modeling will depend on the quality of the data. By carefully cleaning, transforming, and selecting the data, we can ensure that the resulting models and insights are accurate and reliable.

5.1. Data Cleaning

Data cleaning and validation techniques help determine and solve inconsistencies, outliers, anomalies, incomplete data, etc. Clean data helps to find valuable patterns and information and ignores irrelevant values in the datasets. It is very much essential to build high-quality models, and missing or incomplete data is one of the best examples of poor data. In our project, we went through different steps to come up with a clean database.

5.1.1. Step 1 : Deleting unnecessary features

First, we started by deleting the features that won't be useful to us to achieve the goals of our project. In fact, since we are mainly aiming at extracting information about the skills required in each field of study, to develop the most accurate and complete curriculum, some of the features that are in our primary dataset won't help us when we will start the modeling phase. Therefore, deleting them is recommended.
These are the reasons that led us to delete each one of these columns:

- externalApplyLink : We won't apply to any kind of job offer in our project. Thus, we won't need the link to the application page.
- id : We decided to use the number of the row in the dataset to identify each job offer, rather than a random string of characters generated by the "indeed" website.
- postedAt : Usually, the skills required by a specific job position won't be very different in the short term. Thus, the job offers we scraped should provide us with skills that are still in demand at this exact moment, even if the job offer was published a few weeks ago. That's why, knowing when exactly the job offer was published won't be necessary for now.
- scrapedAt : Knowing at what time we scraped our data is irrelevant to us.
- url : We already retrieved all the information that we need in our dataset. We won't need the link to the job offer for the rest of the project.
- jobType : All of the values in this feature are missing.
- companyLogo : The company's name is sufficient.

At the end, the remaining 8 features are: "company", "description", "location", "positionName", "rating", "reviewsCount", "salary" and "field".

5.1.2. Step 2 : Dealing with missing values

Since missing data always reduces prediction accuracy and performance of the model, data must be cleaned and validated through various imputation tools to fill missing fields with statistically relevant substitutes. In our case, when we visualized our data (as the above bar charts shows), we found out that almost a third of the "rating" and "reviewsCount" features were missing. Thus, we need to rectify this situation using a specific strategy. However, before making any decision, we need to understand the type of our missing values.

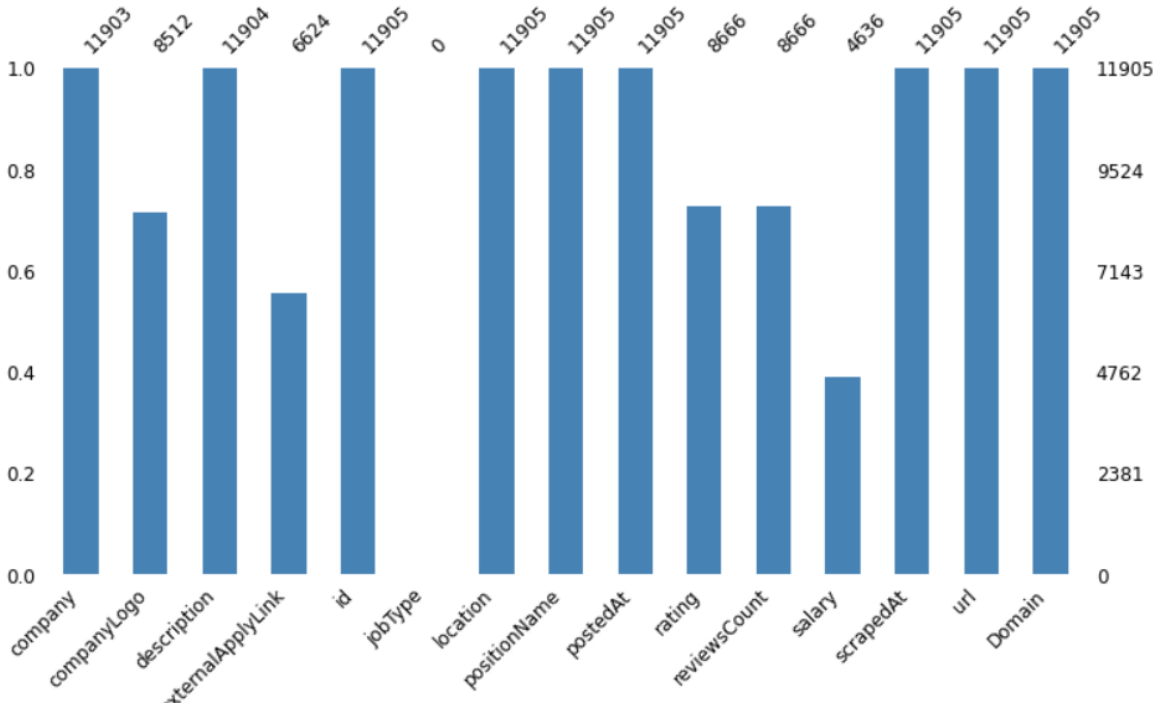

Figure 2: The missing values in our Database (Bar chart)

There are 3 main types of missing values:

- Missing completely at random (MCAR) : This type of missingness occurs when the missing values are unrelated to any other variables or outcomes. For example, if a survey respondent accidentally skips a question or their response is lost due to a technical error, this would be considered MCAR.
- Missing at random (MAR) : This type of missingness occurs when the missing values are related to other observed variables or outcomes, but not to the missing values themselves. For example, in a study on income, participants who have higher income may be more likely to skip questions about income due to a sense of privacy.
- Missing not at random (MNAR) : This type of missingness occurs when the missing values are related to the missing values themselves. For example, a survey question about a sensitive topic, may not be answered by participants who fear discrimination or stigma based on their response, would be considered MNAR. MNAR can be particularly problematic because the missingness is related to the variable being studied, which can bias the results if not accounted for appropriately.

In our case, the missing values of the "rating" and "reviewsCount" features are most likely due to some user behavior. Generally, people don't feel the need to rate an article (in this case a job offer) that they just read. Therefore, the missingness is likely to be Missing not at random (MNAR). In this situation, we should consider using a more advanced method for imputation.
After various research on the subject, we decided to go with the "Multiple Imputation" (MI) method, using the fancyimpute library. The basic idea behind this approach is to impute the missing values multiple times using a statistical model and then combine the results to produce a single estimate. By creating multiple imputed datasets and analyzing each one separately, MI accounts for the uncertainty in the imputed values and produces more accurate estimates of the missing data, compared to single imputation methods.

Also, since we are planning on using the "salary" column in the future, we need to deal with its missing values too. This case is a bit particular because more than half of its values are missing. This is probably due to the fact that most companies prefer to discuss the salary directly with the concerned person, instead of adding it in the job offer description. For this column, we decided to extract the numeric characters from each observation and then calculate the average yearly salary of each specialty and use it to impute those missing values.

5.1.3. Step 3 : Dealing with duplicated observations

Now that we don't have any missing value left, we need to get rid of the duplicated observations. Most companies will publish a job offer multiple times. The most important feature "description" will thereby be duplicated which can lead to a biased data. That's why we decided to drop all the duplicate rows, except the first one.
At the end, we will have 9700 remaining observations in our database.

5.1.4. Step 4 : Dealing with the outliers

When visualizing the outliers of the numeric features (rating, salary and reviewsCount), we noticed that there wasn't a real outlier :

- The numbers in the "rating" feature fluctuate between 1 and 5
- Each job offer should have its own number of reviews (which is calculated automatically by the website. The value cannot be wrong)
- Each company will suggest a salary for the job position it is offering. There is no illogical salary in our dataset

There is no need to make any kind of treatment on the outliers in our case.

5.2. Feature Transformation

5.2.1. Feature Encoding

Feature encoding is the process of transforming categorical or text data into numerical data that can be used in machine learning models. Machine learning algorithms typically require numerical data as input, but many datasets contain categorical features that are represented

as text or strings. Feature encoding allows these categorical features to be transformed into numerical data so that they can be used in machine learning models.

At the beginning, all 8 of our features were objects. Throughout our work, we converted the “rating”, “salary” and “reviewsCount” features to numeric values. Also, we noticed that the “location” feature contains only two different values, which are: “US” and “Europe”. So, we converted it to binary using the Label Encoding method.

5.2.2. Normalization

Normalization is the process of scaling numerical data in a dataset to a coon range or standard scale. This is done to avoid bias towards features with larger scales or ranges, which can have a significant impact on some machine learning algorithms. Normalization can be performed in various ways, but one of the most common techniques is min-max scaling, which rescales all feature values to a range between 0 and 1. This is done by subtracting the minimum value of the feature from each value and dividing the result by the range of the feature. Normalization is important because many machine learning algorithms assume that the features are on a similar scale and can be treated equally.

After encoding some of the features of our dataset, we obtained 4 numeric variables. We then applied the normalization process on them, to ensure that each feature contributes equally to the analysis, resulting in more accurate and robust machine learning models.

5.3. Feature engineering

The most important feature in our dataset, that will help us generate the perfect curriculum, is definitely the “description” one. It contains all the required skills to each job offer. In order to get the most of it, we need to extract the keywords. In this section of the project, we will use the NLP (Natural Language Processing) approach to analyze and understand the meaning of the text, as well as the structure and grammar of the language. This technique will help us extract the keywords to simplify the classification of the competencies needed in each field nowadays.

Figure 3 : Visualization of the keywords in the "description" feature

First, we used the “stopwords” functionality of the NLTK (Natural Language ToolKit). This is used to clean up all the description paragraph by removing the stopwords (such as the pronouns, the prepositions, the auxiliaries, etc.) and the punctuations (such as “!”, “$”, “%”, “&”, “\”, “(“, “)”, “*”, etc.). We only left the “+” and the “#” symbols because they are used when referring to the C++ and C# programming languages. We even lowercased everything to avoid any kind of case sensitivity issue.

Second, in order to enrich our skills’ list, we did further research on our own to add new in demand soft and hard skills. These will be used to identify the skills needed for each job nowadays and the ones that are common to all of the specialties.

Last but not least, we classified all the skills that we have collected from the lists of subjects of all specialties. Then we added new columns of subjects to ensure a better classification of the skills contained in the 'description' feature for each subject.

Since we are also planning on adding a resume evaluation skill, we used OCR to extract the skills that are specified in the resume file.

5.4. Statistical Analysis

In our initial research phase, we embarked on a comprehensive exploration of study plans and subjects, aiming to understand the specific skills required for each subject. Based on our research, we were able to identify a range of 70 distinct subjects. Then we analyzed the job “description” feature of our data, and we were able to extract valuable information regarding the skills sought by employers in various industries and sectors. This information served as a foundation for our analysis and allowed us to create an extensive list of skills that are highly relevant and in demand in today's job market. Using this list, we were then able to formulate study plans based on the identified skills.

The next step was dedicated to focusing on extracting the modules that were common across different study plans. For that, we used the “description” feature to extract the skills and the subjects that are common to most of the job offers. By analyzing the subjects and their corresponding skills, we were able to identify similarities and overlaps among them. This process allowed us to establish a strong correspondence between subjects and the shared skills they contained.

Some of the common subjects that we extracted are: data mining, cloud, web service, web frontend, backend, mathematics …

Those common modules were then deleted from our list of subjects since they will 100% be taught no matter the specialty. They were also deleted from the “description” feature of the job offers. Each option is now left with a list of subjects that are only specific to it, and each job offer is now left with a list of the most advanced required skills. We even discovered that some of the job offers only required the skills taught in the common modules. We define them as “Typical job offers”. In that case, the job “description” feature became empty. Hence, to make sure that it doesn’t alter the results of the training in the modeling phase, we decided to delete those observations from our dataset.

Finally, we focused on identifying the top five modules for each specific option. In other words, we wanted to answer the following question: What are the skills that would provide the option with a real added value and ensure its alignment with the current job market requirements? To realize that, we analyzed the observations grouped by each field of work (for example Data Science job offers only) and we determined the skills that appeared most frequently within each option. This helped us prioritize their importance. For example, as a result for the Data Science field, we found out that the “Artificial Intelligence” skill was one of the most in demand skill. Hence, teaching it would be a wise decision for the universities. This statistical approach provided us with an order of priority that was used for selecting option-specific modules after eliminating the common ones.

As a result, we discovered promising outcomes that paved the way for identifying 2 curricula. One is accredited by the CTI (Commission des Titres d’Ingénieur) and the other one is not accredited by the CTI.

6. MODELING & EVALUATION

The modeling phase in the CRISP-DM methodology is a pivotal stage. This is the point at which our hard work begins to pay off. The data we spent time preparing is going to be transformed into predictive or descriptive models. It involves selecting and configuring modeling techniques, training and evaluating models, and iteratively refining them for optimal performance. The objective is to develop models that accurately predict outcomes and classify data. Evaluation metrics and techniques are used to assess model performance, and iterative refinement processes are employed to enhance the models. We need to look more broadly at which model best meets the business needs. The modeling phase plays a crucial role in leveraging data to drive informed decision-making and achieve data mining goals. We decided to apply machine learning methods, specifically classification models, to predict whether a subject should be taught or not, define the popular competencies and recognize job market patterns.

The performance of a model significantly depends on the value of hyperparameters. Doing several combinations manually could take a considerable amount of time and resources and thus we decided to use the GridSearchCV algorithm to automate the tuning of hyperparameters. GridSearch is used to find the optimal hyperparameters of a model which results in the most 'accurate' predictions. It tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function, we get accuracy/loss for every combination of hyperparameters, and we can choose the one with the best performance and the lower error score.

6.1. Random Forest Model

6.1.1. Definition

Random forest is a commonly used machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. It was trademarked by Leo Breiman and Adele Cutler.

The fundamental concept behind random forest is a simple but powerful one: The Wisdom of Crowds. In data science speak, the reason that the random forest model works so well is that: *A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.* In other words, it is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Ensemble learning uses two types of methods:

- Bagging: It creates a different training subset from sample training data with replacement & the final output is based on majority voting.
- Boosting: It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy.

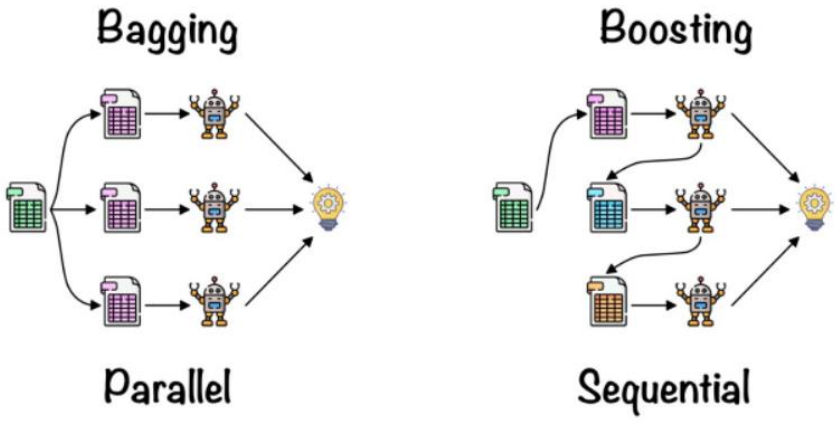

Figure 4 : The Bagging & Boosting ensemble learning methods

Random forest works on the Bagging principle. consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. In other words, it combines the output of multiple decision trees to reach a single result. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. These are the steps involved in Random Forest Algorithm:

Step 1: In the Random Forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on *Majority Voting or Averaging* for Classification and regression, respectively.

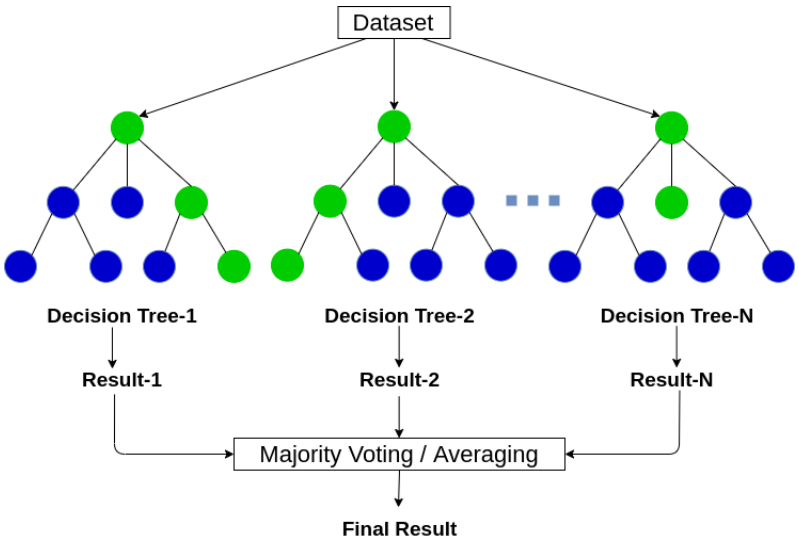

Figure 5: The Random Forest Model

6.1.2. Modeling

These are the best hyperparameters obtained after using the GridSearch algorithm on the Random Forest model:

- n_estimators = 300
- random_state = 42

6.2. K-nearest Neighbors Model

6.2.1. Definition

K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification problems. It assumes that similar things exist in close proximity. In other words, similar points can be found near one another. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closest to the test data.

Suppose there are two categories, Category A and Category B, and we have a new data point x1. The question that we need to answer is: in which category will the data point x1 lie. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.
These are the steps involved in the KNN Algorithm:

Step 1: Select the number K of the neighbors.

Step 2: Calculate the Euclidean distance (distance between two points) of K number of neighbors.

Step 3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step 4: Among these k neighbors, count the number of the data points in each category.

Step 5: Assign the new data points to that category for which the number of the neighbor is maximum.

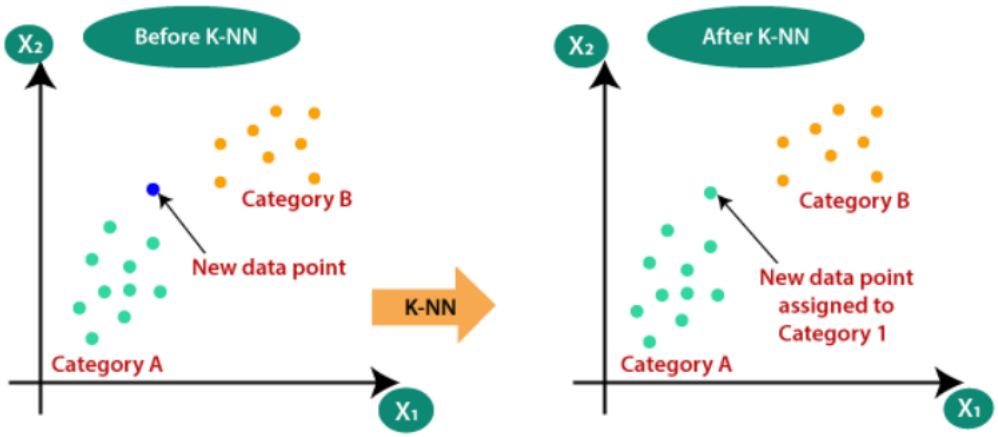

Figure 6: The K-nearest Neighbors Model

6.2.2. Modeling

These are the best hyperparameters obtained after using the GridSearch algorithm on the K-nearest Neighors model:

- n_neighbors = 20
- p = 2
- weights = 'distance'

6.3. Support Vector Machine Model

6.3.1. Definition

Support Vector Machine (SVM) is a supervised learning machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems, such as text classification. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. We choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane/hard margin. One more interesting information about the SVM algorithm: it has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers.

Suppose there are two categories, Blue and Green. Our goal is to build a classifier that can classify unknown points into one of these classes according to their features. The obvious way of dividing these classes would be to draw a line that separates the entries with “blue” and “green” classes. But there is an infinite number of possible lines to draw. How can we know which one is correct?

These are the steps involved in the SVM Algorithm:
Step 1: We only pay attention to the points on the boundary between two classes and can safely forget about many others.
Step 2: Perform a calculation. We linearly connect each pair of points on opposite sides, and the center of this line is equidistant from the classes on the boundary lines. These support points on the boundary lines are actually the sequence containers representing arrays. From now on, we will call them support vectors.
Step 3: The line formed based on these middle points is our boundary line. It can be a straight or a curved line.

Step 4: The boundary line that we found is going to helps us classify points to each of these classes. The points that fall on the “green side” of the line will be classified as green, the points that fall on the “blue side” will be classified as blue. For this reason, we call it a decision boundary.

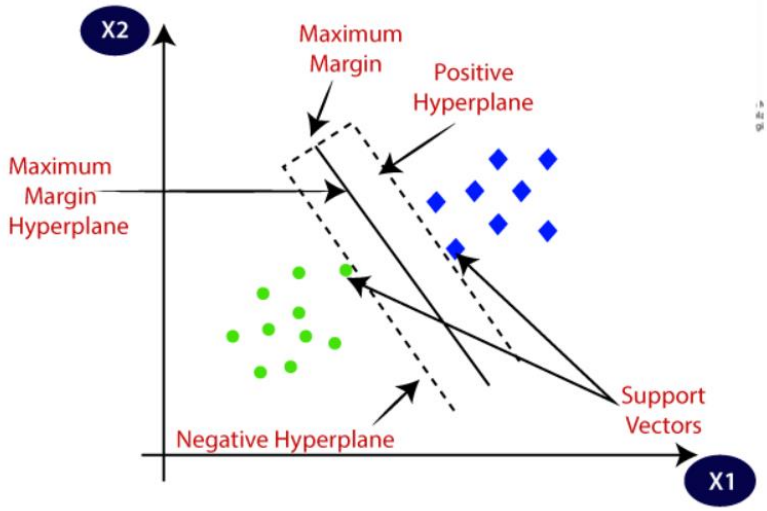

Figure 7: The Support Vector Machine Model

6.3.2. Modeling

These are the best hyperparameters obtained after using the GridSearch algorithm on the Support Vector Machine model:

- decision_function_shape = ‘ovr'
- kernel='linear'

6.4. Evaluation

After training and testing our models, we discovered that certain models performed more accurately for each subject compared to others. For instance, the SVM would perform better for subject X, but the KNN would perform better for another subject Y. Hence, we couldn’t make a real decision and choose one of the models as our reference model. Consequently, we decided to use the three different models: Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (KNN). For each subject, we evaluated the precision of these models and selected the one with the highest accuracy. By considering the precision of each model, we aimed to ensure that subjects were appropriately classified and included in study plans based on their relevance and accuracy. It allowed us to determine whether a subject should be included in a study plan or not based on the chosen model's performance for that specific subject.

7. DEPLOYMENT

A model is not particularly useful unless the customer can access its results. Thus, after months of development and rigorous testing, it is time to deploy our competency-based curriculum platform and make it accessible to students and universities. The deployment phase marks the transition from a development environment to a live system that can effectively serve its intended users. The deployment phase represents the culmination of our efforts in developing the competency-based curriculum platform.

7.1. Technologies used

To ensure a smooth deployment process, several key considerations need to be addressed.

First of all, we decided to use the framework Django to deploy our platform. It is a high-level web framework for building robust and scalable web applications using the Python programming language. It is an excellent choice for web application deployment due to its scalability, rapid development capabilities, security features, strong community support, extensive documentation, versatility, and testing/debugging tools. It provides a robust foundation for building and deploying web applications efficiently and effectively.

Second, we used Docker to simplify the development, deployment and scaling of our application. Docker is an open-source platform that enables developers to build, deploy, run, update and manage containers—standardized, executable components that combine application source code with the operating system (OS) libraries and dependencies required to run that code in any environment. Containers simplify development and delivery of distributed applications.

7.2. Final Solution

We decided to name our platform: Learn2Lead. As it suggests, we are aiming at empowering individuals, especially students, to become leaders in their fields through learning and skill development. The use of "Learn" in the name emphasizes the importance of continuous learning and personal development to stay up to date. Whereas the use of "Lead" in the name signifies the ultimate goal of the platform: to equip individuals with the knowledge and competencies required to lead, make an impact and drive positive change in their professional lives. Learning is the pathway to success and building one’s dream career.

In our platform, we have implemented a feature for universities to provide a comprehensive and tailored curriculum that is in line with current job market demands. This functionality is designed to help universities equip their students with the necessary skills and knowledge that will enhance their employability. By collaborating with educational experts, our platform ensures that universities are offering a curriculum that is up-to-date and relevant, providing students with the best chance of success in their chosen career paths. With this feature, we aim to bridge the gap between academia and the job market, ultimately creating a more efficient and effective transition for graduates into the workforce.

Furthermore, we have also implemented a resume rating system for students which provides a percentage score based on the current job market needs. It can also give suggestions for acquiring the skills required to secure a position in the job market. This feature helps students identify the gaps in their resumes and motivates them to enhance their skills to become more competitive in the job market. With this tool, our platform aims to provide students with a more personalized and insightful job search experience, ultimately increasing their chances of success in their future careers.

Additionally, we have integrated a Power Bi Dashboard in the Backend (only seen by the administrator). This will help us have a constant overview on the evolution of our platform.

Finally, Post-deployment, ongoing monitoring and maintenance will be essential to ensure the platform's stability, security, and performance. Regular updates and patches will be applied to address any identified issues or vulnerabilities. Moreover, we added a review section in our platform. That way, users' feedback will be actively collected to drive continuous improvement and refinement of the platform.

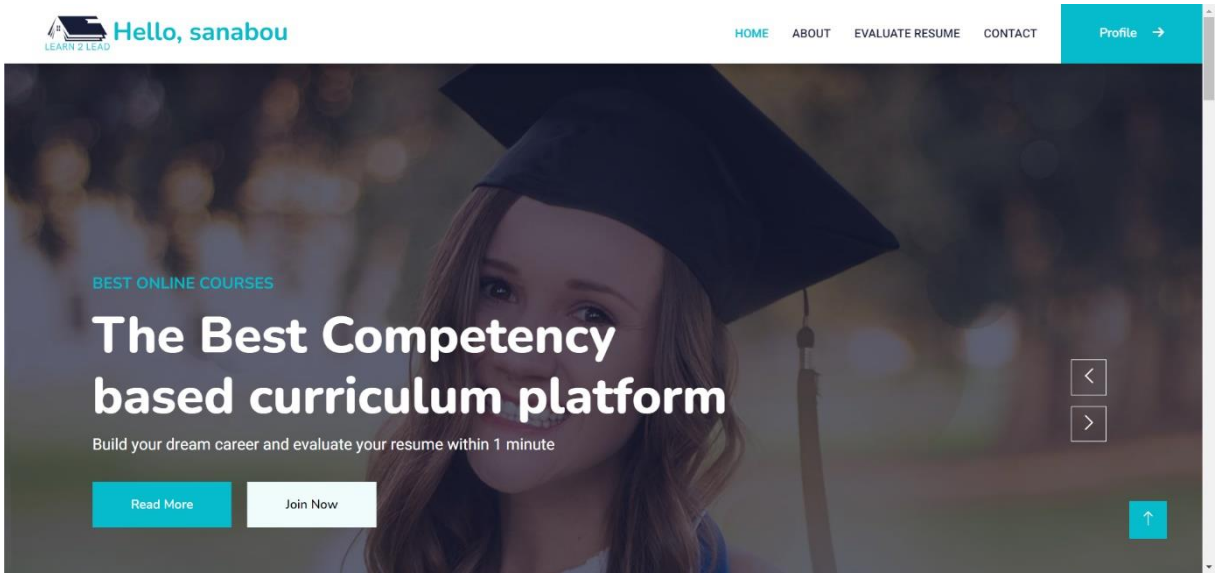

Figure 8: Sneak peek of the Learn2Lead platform

CONCLUSION

Our data science project focused on developing a competency-based curriculum platform with the goal of facilitating students' employability and aligning their skills with the needs of the job market. Through months of dedicated work, our team successfully built the platform from scratch, integrating key features that address the challenges faced by students during their job search.

The platform's primary objective is to provide students with a comprehensive and tailored curriculum that their university can adopt. By utilizing data scraping techniques from the Indeed job platform, we collected a vast array of job offers, which served as a basis for the generation of the curricula. Through data preparation, including statistical analysis, we defined the necessary skills sought by employers. We aim to enhance students' chances of getting hired easily by aligning the curriculum with the actual needs of the job market. This alignment ensures that the skills they acquire throughout their educational journey are relevant and in demand in the current job market. Machine learning algorithms, specifically K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM), were trained to evaluate resumes and provide a score to quantify the match between a student's skills and the job market requirements. Notably, SVM demonstrated superior performance in this task.

Moreover, we implemented a resume evaluation feature within the platform. In order to achieve that, we incorporated Optical Character Recognition (OCR) techniques to extract skills from uploaded resumes automatically. This streamlined the evaluation process and provided valuable insights to students regarding skills that could enhance their resumes further. By suggesting these skills, we aimed to empower students with targeted recommendations for skill improvement, thereby increasing their employability prospects.

Our data science project represents a significant contribution to bridging the gap between education and the job market. By providing students with a competency-based curriculum and a resume evaluation tool, we empower them to develop the skills necessary to succeed in their desired careers. The platform's data-driven approach and use of machine learning techniques ensure the accuracy and relevance of the evaluations and recommendations.

As we conclude this project, we recognize the potential impact our competency-based curriculum platform can have on students' future employability. Our project showcases the power of data science and its potential to revolutionize the educational sector. We believe that our platform can empower universities and students and prepare them for a smoother transition from education to employment. We hope that our platform will contribute to the continued growth and development of students.

REFERENCES

https://knowledgeworks.org/get-inspired/personalized-learning-101/competency-based-versus-traditional/

https://www.linkedin.com/pulse/chapter-1-introduction-crisp-dm-framework-data-science-anshul-roy/

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4

https://serokell.io/blog/support-vector-machine-algorithm

https://www.javatpoint.com/machine-learning-algorithms

https://neptune.ai/blog/building-deep-learning-based-ocr-model

https://www.ibm.com/topics/docker