Machine Learning and Data Science Awareness and Experience in Vocational Education and Training High School Students

Stefan Zlatinov[1], Branislav Gerazov[1], Gorjan Nadzinski[1], Tomislav Kartalov[1],
Igor Atanasov[2], Jelena Horstmann[3], Uroš Sterle[4], and Matjaž Gams[5]
[1]*Faculty of Electrical Engineering and Information Technologies*
Ss Cyril and Methodius University in Skopje, Macedonia
[2]*Secondary municipal vocational school "Ilinden"*
[3]*Electrical Engineering School "Mihajlo Pupin" in Novi Sad, Serbia*
[4]*School centre Kranj, Kranj, Slovenia*
[5]*Jožef Stefan Institute, Ljubljana, Slovenia*
zlatinov@feit.ukim.edu.mk, gerazov@feit.ukim.edu.mk, gorjan@feit.ukim.edu.mk, kartalov@feit.ukim.edu.mk

***Abstract*—Data Science Machine learning and are increasingly important in the world's economy and there is an increasing gap in the job market of skilled workers. To address this the Valence project proposes the design and implementation of a Data Science and Machine Learning focused curriculum in VET high schools. As a precursor to this, we designed a survey to assess the awareness/experience in these areas in students in Vocational Education and Training high schools. The survey was distributed across the three partner VET institutions. The analysis shows that While most students are aware of these two areas, only a small proportion of them have any practical experience in them or have followed an online tutorial. This reaffirms the need for the design and deployment of an accessible Data Science and Machine Learning curriculum.**

***Index Terms*—machine learning; data science; high-school education; vocational education; survey**

I. Introduction

The total amount of digital data that we create/generate each day is growing exponentially. According to estimates, to the end of 2021, there will be 74 zettabytes of generated data in the world [1] . That is expected to double by the end of 2024. This huge amount of data can be exploited by many industries and open new business opportunities. Indeed, according to the final study of the European Data Market tool, the value of the Data Economy exceeded the threshold of 400 Billion Euro in 2019 for the EU27 plus the UK, with a growth of 7.6% over the previous year [2]. This growth is complemented by an increase in the number of data professionals reaching 76 million in 2019, which is 3.6% of the total workforce - an increase of 5.5% over the previous year. The increased need of data professionals has led to an imbalance between the demand and the supply of data skills in Europe with a gap of approximately 459,000 unfilled positions corresponding to 5.7% of total demand. The data skills gap is forecast to continue as demand will continue to outpace supply [2]. This is aggravated by the reported lack of analytical skills as a key challenge by 43% of employers [3].

To answer this imbalance, there is an urgent need for more widespread education of new generations of students in the disciplines of Data Science (DS) and Machine Learning (ML). At the University level, nearly every technical university has a series of courses dealing with data science and machine learning. However, there is also an increasing trend of having introductory courses in data science and machine learning at non-technical universities. One example is the introductory course for Data Science proposed by [4], implemented at Harvard College and in the School of Public Health, Boston, MA on a diverse group of students in terms of their knowledge of programming and statistics. Moreover, Data Science and Machine Learning are increasingly becoming an integral part of education at the elementary and high school levels. A comprehensive analysis of some 30 instructional units across different schools and curricula, shows a rising trend of courses covering machine learning per year, that is essentially skyrocketing since 2018 [5]. Some of these have explicitly addressed the need for introducing data science in Vocational Education and Training (VET) [6].

The Erasmus+ KA202 project VALENCE - Advancing machine learning in vocational education is focused on developing a curriculum and an integrated free and open-source software platform for teaching Machine Learning and Data Science.[1] The primary target audience are students attending VET high-schools, but the modularity of the platform will allow its wider usage. The project will deploy and test the curriculum in the three partner institutions: the Kranj School Centre in Kranj, Slovenia, the Electrical Engineering School "Mihajlo Pupin" in Novi Sad, Serbia, and the Vocational High

[1] https://valence.feit.ukim.edu.mk/

School “Ilinden” in Skopje, Macedonia.

To assess the awareness of and experience with Data Science and Machine Learning, we designed an online survey that was deployed to the students attending these three VET high schools. The analysis of the results shows that although these topics are increasingly present in our daily lives, high school students have only partial awareness of them, and very few have any direct hands-on experience. These survey results will be of great importance in the development of the curriculum and the platform for teaching Machine Learning and Data Science to VET high-school students.

In Sections 2 and 3 of this paper we will present the design and the deployment of the survey, respectively. In Section 4 we will give an in-depth look into the survey results and conduct a thorough analysis of the answers, before using Section 5 to give a conclusion and outline the most important conclusions going forward.

II. Survey Design

The survey was made comprehensive and thorough. It comprises 8 sections:

1) General questions - serve to obtain personal information about the student without eoprdizing their anonimity, and include: age, sex, school, specialization and grade average,
2) Awareness of DS and ML - a series of 6 questions to assess if the student has heard of or used DS or ML, do they grasp the usage potential of DS and ML, and do they know how to define them in their own words,
3) Contact with, and exposure to DS and ML - questions that assess the frequency and continuity of the student’s use of modern IT platforms, including social media, video streaming services, video games and communication apps,
4) Interest in DS and ML - questions to assess the interest of the student in learning DS and ML, as well as their particular applications,
5) Experience with DS and ML - 2 short questions allowing students to express any practical experience they have with DS and ML,
6) General IT and language skills - questions to assess the programming experience of students as well as their proficiency in English,
7) Learning preferences - evaluation of the perceived benefit and personal preference of the use of various teaching methods including: traditional lectures, course books, homework projects, video lectures, online courses, interactive demos, and work on practical projects,
8) Extras - a group of miscellaneous questions on topics that include: cyberbullying, sports, music, movies, art, and languages.

In total the Survey contains 72 questions, 45 of which are questions in Sections 1 - 7 and directly relate to DS and ML and the development of the VALENCE curriculum, and 27 questions are in the Extras section. The latter, although unrelated to DS and ML, will serve as a rich source of practical examples in the process of the design and development of the VALENCE curriculum.[2]

III. Survey Deployment

The initial version of the Survey was designed in English, and after its finalization, it was translated into the languages of the partner VET High Schools: Slovenian, Serbian and Macedonian. Each translation was performed by a native speaker.

The Survey was deployed in mid-May 2021, in the three partner VET high schools to over 1,000 students of all levels and study profiles. There was a slight preference to distribute the Survey to Computer Science students, as they were identified as high performers. The schools were the Kranj School Centre in Kranj, Slovenia, the Electrical Engineering School “Mihajlo Pupin” in Novi Sad, Serbia, and the Vocational High School “Ilinden” in Skopje, Macedonia. The Survey was left open for nearly one month, in order to give more students a chance to participate. Moreover, since this was near the end of the school year, when students are busy with final projects and state exams, as well as preparations for the prom, they were reminded about filling in the Survey several times. At the end a total of 857 students responded to the Survey.

IV. Results

There were a total of 857 participants in the Survey with the majority, 550 participants, coming from Serbia, then 170 from Macedonia, and 137 from Slovenia.

A. Preprocessing

We first preprocessed the data and identified participants that gave one or more inadequate answers to the questions with textual response. For example, many wrote jokes about their peers, few of them spammed the survey with arabic or chinese letters, while one pupil was determined enough to fill many fields with more than 5000 characters. A lot of them exploited the other option of the sex questio in order to express their creativity. The participants identified using this criterion were eliminated from further analysis. Although seemingly strict, i.e. having one senseless answer discard all the other 72, we decided that this is necessary to maintain stronger validity of the overall analysis of the results. Nearly 25% of the participants were discarded in this way.

B. Demographic distribution

Figs. 1 – 3 show the demographic distribution of the students, i.e. age, sex and study profile, separated by high school (country). We can see that the students are aged 14 – 20, but most are 15 – 17 years old. They are predominantly male, which is not surprising for VET schools. However, we can see that there is a clear difference in the number of women across the three countries – female students make 23.5% of the students in Macedonia, 8.76% in Serbia, and only 2.26% in Slovenia. Regarding their study profiles, almost

[2] The Survey can be found on the following link https://valence.feit.ukim.edu.mk

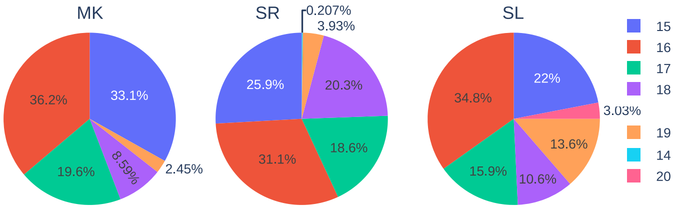

Fig. 1. Age distribution of the participants in the Survey for each country.

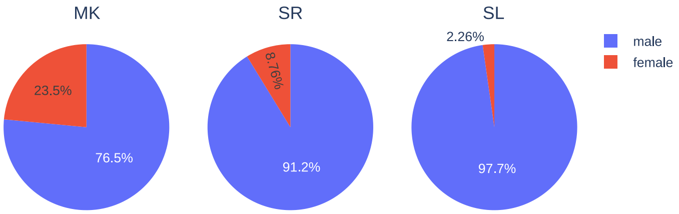

Fig. 2. Sex distribution of the participants in the Survey for each country.

50% of students are Computer Science students, followed by Electronics and Automatics, except in Macedonia where we have a more balanced distribution. This is in line with our preferences in distributing the Survey, but also reflects the number of students in the different study profiles.

C. Awareness of DS and ML

We gauge the level of awareness of DS and ML through asking "Where did you hear about … ?" for DS and ML separately. Fig. 4 shows the analysis of the results. We can see that, even though these subjects are not part of the curriculum, most students have already heard about their existence, with ML being the more familiar term. This reflects their omnipresence in the high tech world we live in today. Despite these encouraging results, some 40% of the students have not yet heard about ML or DS, validating the need to have them included in the curriculum. As expected, the internet is the dominant source of information, and precedes the other sources of information in the chart.

When asked about the definition of the terms DS and ML, two patterns emerge in the student's definitions. The first one is the classic "Data science is the science about data". The second pattern, broadly exploited by the Macedonian pupils, is the first paragraph of the respected Wikipedia page: "Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains."[3]

D. Experience with DS and ML

Delving deeper, we asked the students whether they have used DS or ML in a personal project. Only a handful of Serbian students responded affirmatively with the top three listed answers being neural networks, chatbots and image processing. Fig. 5 shows how many students followed an online tutorial about DS or ML. We can see that Slovenian students are in the lead with Serbian students in second place. The most interesting topics of the tutorials were general programming and ML interest for the Serbian and Slovenian students, while robotics and ML application were the ost interesting for the Macedonian students. However, upon closer analysis, almost a half of the provided responses clearly show that the students are still largely unfamiliar with what constitutes DS and ML, and have a hard time drawing the line between them and classical engineering.

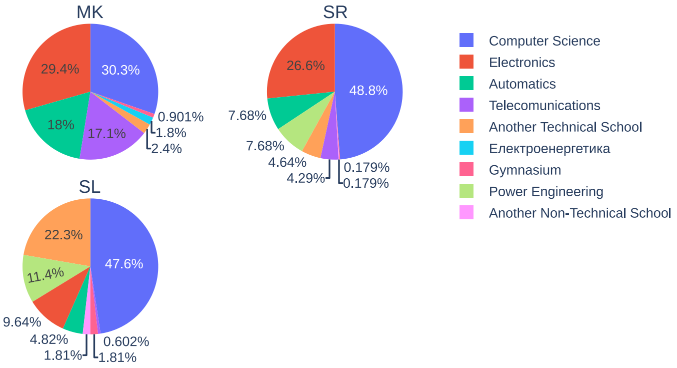

Fig. 3. Study profile distribution of the participants in the Survey for each country.

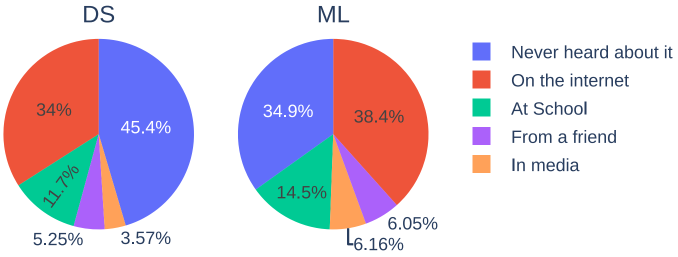

Fig. 4. Awareness of DS and ML of the participants in the Survey.

E. Readiness for DS and ML

To indirectly evaluate the readiness for practical work in DS and ML we asked students whether they have any experience writing code in Python. Fig. 6 shows that the Slovenian students are ahead in Python programming experience, followed by the students in Serbia and Macedonia. In contrast, the use of Jupyter Notebooks is level at 5% in all of the three high schools. Speaking of general programming experience, Slovenian and Serbian students are much more versatile with many of the students ticking two or more programming language checkboxes. On the other hand, Macedonian students have exclusively and in large numbers marked themselves as knowledgeable in C++, reflecting the focus on this programming language in the Macedonian curriculum.

F. Interest in DS and ML

Next, the students were asked to express their curiosity about learning DS, ML, and statistics. Fig. 7 shows that interest in these areas is varied. The mean expressed interest,

[3] https://en.wikipedia.org/wiki/Data_science

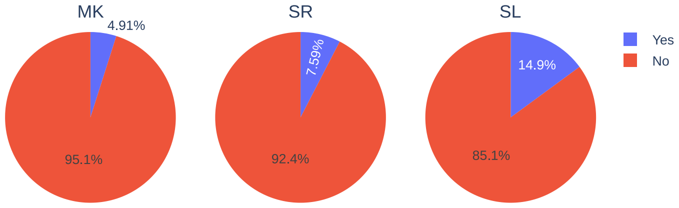

Fig. 5. Ratio of students that followed a DS or ML related online tutorial for each country.

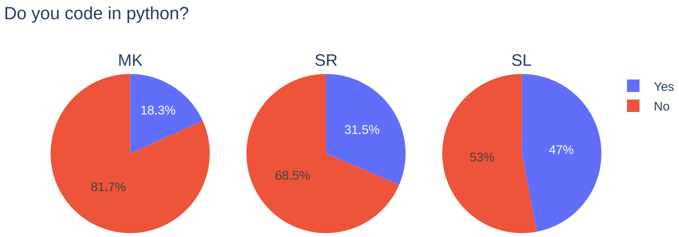

Fig. 6. Python experience of the participants in the Survey for each country.

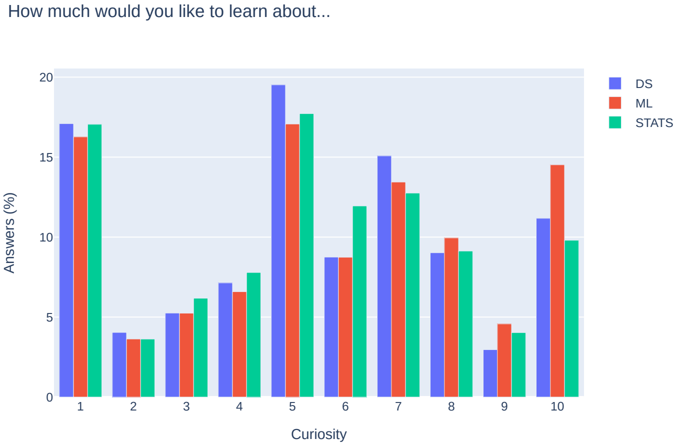

Fig. 7. Interest in learning Data Science, Machine Learning and Statistics.

TABLE I
Statistics of the interest in learning Data Science, Machine Learning and Statistics.

	Data Science	Machine Learning	Statistics	Average
mean	4.35	4.63	4.3	4.43
median	4	5	4	4.33

as can be seen in Table I, is 4.4/10, and differs between the three areas. ML is the most interesting with more than 50% of students expressing interest at 5 and above, while DS and Statistics are slightly less interesting. In Fig. 7 we can see that the largest group of students, almost 20%, expressed themselves as neither being interested nor disinterested about learning DS and ML choosing 5. There are also pronounced groups giving positive scores of 7 and 10. It should be noted however, that even though there are a large number of positive responses, there is a significant portion of students (around 17%) that would not want DS, ML or Statistics included in their curricula at all. This is rather discouraging, but might correlate with the students that feel negative about the education process in general.

We also queried their particular interest in the most popular applications of DS and ML. The results show that the students exhibit a huge interest in ML applied to robotics, followed by image processing and speech recognition. On the other end, w.r.t. DS applications, the majority of the answers are uniformly distributed between business intelligence, digital advertising, and internet search.

V. Conclusion

The analysis shows encouraging results about the awareness of the fields of Data Science and Machine Learning of students from three representative VET high schools from Slovenia, Serbia and Macedonia. The number of students participating in the designed Survey, as well as the established quality of their answers, gives a high level of validity of the results. The analysis shows that students don't have ample awareness and experience in the fields of DS and ML even in VET high schools. Even if most students are aware of DS and ML, only a small proportion of them have any practical experience in them or have followed an online tutorial. This reaffirms the need for the design and deployment of an accessible DS and ML curriculum. In fact, their experience with Python, shows that a large number of students already have good prerequisites for following such a course.

VI. Acknowledgement

The work has been financed by Erasmus+ KA202 project VALENCE: Advancing Machine Learning in Vocational Education.

References

[1] A. Holst. (2021) Amount of data created, consumed, and stored 2010 - 2025. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/

[2] European Centre for the Development of Vocational Training, "Learning outcomes approaches in VET curricula: A comparative analysis of nine European countries," https://www.cedefop.europa.eu/files/5506_en.pdf, 2010.

[3] G. Press. (2010) The supply and demand of data scientists: What the surveys say. [Online]. Available: https://www.forbes.com/sites/gilpress/2015/04/30/the-supply-and-demand-of-data-scientists-what-the-surveys-say

[4] S. C. Hicks and R. A. Irizarry, "A guide to teaching data science," *The American Statistician*, vol. 72, no. 4, pp. 382–391, 2018.

[5] L. S. Marques, C. Gresse von Wangenheim, and J. C. Hauck, "Teaching machine learning in school: A systematic mapping of the state of the art," *Informatics in Education*, vol. 19, no. 2, pp. 283–321, 2020.

[6] SEnDIng Erasmus+ project. (2017) D2.3: Vocational curricula/educational modules for Data Science and Internet of Things VET program. [Online]. Available: http://sending-project.eu/attachments/article/71/SEnDINg_DLV2.3-1st_version.pdf