Process of building a dataset and classification of VARK learning styles with machine learning and predictive analytics models

D. KRISHNAMOORTHY¹, DUVVURU LOKESH²

¹Assistant Professor, Saveetha School of Management, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India.
²MBA Student, Saveetha School of Management, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India.
Email ID: krishnamoorthy.ds.mba@saveetha.com, damnloku@gmail.com

Abstract: As there is a rise in the online and customized learning platforms, learning style preferences give us insight into better utilization of educational resources available. VARK learning styles are developed by Fleming and Bonwell on the premises of preferred intake of information by the students. VARK model describes four sensory modalities Visual, Kinesthetic, Read/Write, and Auditory respectively for the input. Initially, learning styles are calculated using the VARK questionnaire as an instrument. In this case, machine learning and predictive analytics can help classify learning style by including all the descriptors including demographic and behavioural. This study will explore the data mining process from the raw data collected from the college students through a questionnaire which is an integral part before the analysis to determine possibility to use different descriptors. This study primarily aims to classify the students based on VARK learning styles based on their sensory modalities and build a predictive model using the variables that can influence the learning styles. This study explored the relationship between demographic factors like school and place people grew up. Results proved to contradict those factors. We concluded that with the growth of big data learning style classification, a blend of model algorithms or stacked algorithms like voting classifier can be used to adapt to a user application.

Keywords: VARK learning styles; Machine learning; Entrepreneurship; Data mining; predictive analytics; supervised learning

INTRODUCTION

As the education industry moves into the new era of online learning, it is beneficial for both the students and instructors to classify students based on their learning styles as proposed in the (“How-Do-I-Learn-Best-Sample.pdf,” n.d.). They explain learning style is a term of reference to an individual's preferred way of gathering, organizing and thinking about the information provided. VARK model is the preferred way to examine the individual's preference on taking in and giving out the information. Learning styles are different ways of learning. They include teaching and learning methods, unique to each individual that allows her/him to grasp more information. The idea of learning styles came into prominence in the 1970s and it has been popular study by many social scientists and academicians.

Learning styles are affected by numerous variables, for example, individual experience, various insights and character factors, such as an inclination for learning alone or in a group. Our learning style will impact how we adapt to ordinary assignments throughout our life, for example, using a guide or preparing a dinner. A valuable guide to help comprehend this idea better is the means by which we figure out how to utilize another bit of technology. We can move toward it either by sitting alone, perusing directions from start to finish previously or take a 'hands on' approach like squeezing the various controls to find through experimentation or learn by observing others utilizing the equivalent. This model assists with reflecting about how learning inclinations shift among people. This said however, conditions may likewise decide how every individual discovers some new information. Such models help us to consider how we have inclinations for the way in which we learn. Hence, understanding learning styles approaches helps us to think about a person's prevailing or favored perspective thus helping us to learn better in a lesser time as proposed by (Sreenidhi and Helena, n.d.).

No student or instructor is limited to only one mode for communication input and output. Even in this way, it is fascinating to take note of that there are some prevailing inclinations and a few voids among different students and instructors. A few students and teachers lean toward not just a solid inclination for one specific mode yet additionally relative shortcomings in different modes. For taking in our surroundings, we have a tendency to use our senses - sight, hearing, taste, bit and smell. In educational learning we have a tendency to typically use our sight, our speech and our hearing with less importance placed on style, bit and smell. Some learners like to use all...
their senses at once by experiencing their learning and this comes under kinesthetic preferences (“How-Do-I-Learn-Best-Sample.pdf,” n.d.). As illustrated in (“How-Do-I-Learn-Best-Sample.pdf,” n.d.), the importance of the VARK model is such that both students and teachers can utilize it to improve learning experience and teaching experience respectively. The acronym of VARK stands for Visual, Auditory, Read/Write and Kinesthetic. These are the four sensory modalities which help a learner to learn information. For instructors preferred sensory modalities of the students help them to adapt in the classroom or any platform used for teaching. In order to classify the large number of students based on their learning styles there are excellent possibilities by leveraging the growing information technology especially predictive models from machine learning and deep learning. By using the predictive models there is possibility to build a user interface to classify learning styles.

Adaptive learning systems may use this knowledge to offer more precise personalization by identifying the learning patterns of students, leading to increased satisfaction and decreased learning time. In addition, students can directly benefit from a more precise recognition of learning styles, by being able to exploit their strengths in terms of learning styles, and by recognizing their limitations. In addition, teachers may use this knowledge about the learning style to offer more detailed guidance to students, which, as described before, becomes more useful for students as the recognition of the learning style becomes more detailed as well. In addition, in the same classroom, students with common learning styles will work together to enhance their learning experience and support the teachers with their techniques. Additionally, other stakeholders in the education ecosystem, such as teachers, administrators and parents, can make use of such an approach to improve education in general as proposed (Gomede et al., 2020).

According to (Gomede et al., 2020), computational variables related to the learning style can be adapted from various sources such as questionnaires, databases and registers of the educational institutions. Moreover, these variables give us the source of input of information for example diagrams, workshops, textbooks and recordings. Then the output is extracted as a type of learning style such as Visual, Kinesthetic, Read/Write and Auditory respectively for the input. For this purpose, a computer intelligence tool can come in handy for both the students and instructors to determine the learning strategies and teaching methods respectively. As (Li and Abdul Rahman, 2018) points out, although combined detection methods have been successful in the institutions, instructors have to use a lengthy questionnaire to collect the data from the students. This has been a huge drawback as it consumes more time and inefficient as it involves both static and dynamic approaches.

In the favour of the above context, this study will explore the data mining process from the raw data collected from the college students through questionnaire which is an integral part before the analysis to determine possibility to use different descriptors. This study primarily aims on selecting a suitable model that can be adapted by the educational institutions as a user interface. Objectives of this study are to classify the students based on VARK learning styles based on their sensory modalities and build a predictive model using the variables that can influence the learning styles.

Our research idea is based on the rich knowledge acquired by our peer teams across the university.(A.C.Gomathi, S.R.Xavier Rajarathinam, A.Mohammed Sadiq, Rajeshkumar, 2020; Danda et al., 2009; Danda and Ravi, 2011; Dua et al., 2019; Ezhillarasan et al., 2019; Krishnan and Chary, 2015; Manivannan, I., Ranganathan, S., Gopalakannan, S. et al., 2018; Narayanan et al., 2012, 2009; Neelakantan et al., 2013, 2011; Neelakantan and Sharma, 2015; Panchal et al., 2019; Prasanna et al., 2011; Priya S et al., 2009; Rajeshkumar et al., 2019; Ramadurai et al., 2019; Ramakrishnan et al., 2019; Ramesh et al., 2016; Venugopalan et al., 2014)

Currently we are working on “A Study on the process of building a dataset and classification of VARK learning styles with machine learning and predictive analytics models”.

THEORETICAL FRAMEWORK

A student can struggle in performing in the exams if his learning strategies do not align with his learning style or an instructor can fail to make his/her students grasp the information he/she is teaching due improper understanding about the students’ sensory modalities. Both situations need an efficient learning strategy. This personalized learning strategy can be determined through a VARK model as proposed in (“How-Do-I-Learn-Best-Sample.pdf,” n.d.). As (Li and Abdul Rahman, 2018) illustrates the need for a dynamic model to classify the students’ learning style, there is a need for the tool that can interpret personal information and inputs from the student and predict his/her learning style.
As (Gomede et al., 2020) proposed that the automatic process is divided into three stage problems. First collection of data through a questionnaire to prepare a dataset with target variables. Second is to collect the information from the learner's portal as descriptors. And at last train the dataset and make predictions through a user application. As (Li and Abdul Rahman, 2018) illustrates with default learning styles, students join the method. When they communicate with the system, their learning styles will be modified differently based on their learning behaviour.

REVIEW OF LITERATURE

1. (Maseleno et al., 2016) illustrated the multicultural perspective to the learning styles in higher education through diagnostic profiling. They created the data which consists of six sections which are culture, learning preferences, cognitive learning styles, creative skills, motivation and students' background knowledge which can help in performing data analytics.

2. (Švarcová and Jelínková, 2016) presents the results of the pilot portion of research aimed at defining learning styles among selected university students. The views on learning styles vary in many respects, as do the points of view themselves. Nevertheless, most opinions highlight the need for specific students and pupils to respect and accept learning styles. The styles of learning are deeply rooted in the biological conditionality of the individual, so they are not easy to alter. They conclude stating that the pieces of knowledge need to be tailored to the teaching process so that the learning outcomes are as successful as possible.

3. (Li and Abdul Rahman, 2018) proposed a learning style detector using the Bayesian network using tree augmentation. Their experimental results proved that their method is more accurate than the results achieved using the Bayesian network. They proved that, if we take into account problems in the identification of learning styles, the proposed naive Bayesian tree augmented model helps us to discover the learning styles of students in a highly precise way.

4. (Khongpit et al., 2018) conducted a study on classification of the VARK learning styles among the computer science students. They illustrated that most of the computer science students around 53% are multimodal learners. In that most of the students fell in the category of bimodal of V and K. The study outcome shows that in order to enhance their motivation and understanding, the teaching material design should be achieved by designing imaginative activities and atmosphere in accordance with the learning abilities of the learners.

5. (Aissaoui et al., 2019) have suggested an approach to automatically classify the learning style based on the habits of the current learners and using techniques of online use mining and machine learning algorithms. In order to group them into 16 learning style combinations based on the Felder and Silverman learning style model, the captured learners' sequences were provided as an input to the K-modes clustering algorithm. Then, to predict a student's learning style in real time, the naive Bayes classifier was used.

6. (Kolekar et al., 2018) presented an approach to identify Felder and Silverman learning style. Using the Moodle platform, an e-learning application was created with the functionality to collect learners' usage data. The use data is used to group the learners according to FSLSM learning categories. On the portal, customization is created by designing an adaptive user interface for each learner based on the FSLSM learning style. Their portal recognizes the learning style of the students and then offers content and customizes the User Interface (UI) based on that learning style.

7. (Labsi et al., 2019) experimented with the Felder and Solomon’s (1999) Index of Learning Styles (ILS) on the interior design students in Saudi Arabia. They introduce the clustering-based approach where they use agglomerative clustering algorithm and dendrogram to classify the students based on their personal data and preferences in learning styles. Their findings give 57% of students are classified as multiple learning styles.
8. (Viloria et al., 2019) research helps students from various professions to understand the learning style preferences of college students and encourages teachers to go to college. After evaluating the learning preferences of students from different professions, the findings obtained led to the conclusion that college students have greater preferences for the reflector learning style followed by the analytical, pragmatist, and activist types of students from different professions.

9. (Dantas and Cunha, 2020) presents the debate on the link between different learning styles from the kolb’s model to Fleming’s VARK model. They theorized based on their result selecting a particular learning style, in other words, relying on a single form of stimulus for the coordination of learning tasks as a presupposition for better learning, tends to be restrictive. While accepting individuality and the preferences of the subject, the most suitable approach for the construction of learning would be to provide students with various stimuli, similar to the different styles.

10. (Zhang et al., 2020) proposed to define and classify the learning styles of students, by introducing a learning style classification method based on the deep belief network (DBN) for large-scale online education. They illustrate that DBLS provides better accuracy compared to traditional approaches.

11. (Gomede et al., 2020) proposed a deep multi target prediction model that can classify Felder–Silverman learning styles. Their model consists of an artificial neural network for the automatic detection of learning styles. The results obtained by them show that learning styles allow e-learning systems to improve the learning processes of students.

METHODOLOGY
In this study exploratory research is adapted as the nature of the problem statement is to investigate different predictive models which can be adapted for the classification of the learning styles. This research also explores the entire process of building an automatic learning style detection model and to build a dataset for training and testing the learning style detector. Data is collected as per the dataset collection process proposed by (Gomede et al., 2020). In this study data collected is primarily from the full time college students. The same process proposed by (Gomede et al., 2020) are adapted according to the objective of this study. Data consists of three parts primary VARK data, demographic data and student preferences data. The dataset preparation process is explained in the diagram below.

![Diagram of dataset preparation process](image)

**Fig.2: The updated process to build a model for detection of learning style.**

The data collection is done through the snowball sampling. Snowball sampling is where a research participant recruits other participants for a study. It is used where potential participants are hard to find. In this case initially five participants are selected and those five participants are asked to select five participants each and so on. In this study this type of sampling can help include students from all backgrounds and multiple locations which in turn can help build a versatile predictive model.

Questionnaire is the instrument that logs in the data to initiate an experiment. This study uses the questionnaire to capture the classification data as well as the descriptors. As (Gomede et al., 2020) suggested, Questionnaire consists of three parts, primary VARK 8.0 questions, demographic questions and student preferences questions in the form of scenarios.

Demographic variables include name, mail id and college name which corresponds to each student. Then possible demographic descriptors like age, gender, college location, educational designation, educational stream, place grownup during schooling, school region, school type and school board.

VARK 8.0 questionnaire from VARK-Learn Limited is included by altering as needed for this study. This part of the questionnaire is for classification and to form the target variable in the dataset with the student preferences. This part of the data is discarded after classification. The last part of the questionnaire consists of descriptor
variables that give weightage for visual, auditory, read/write and kinesthetic attributes of each learner. This part of the data is taken in the form of Likert scale and assigned weight as per attributes. These weightages serve as numerical descriptors in the dataset for determining the learning style.

**DATA ANALYSIS AND RESULTS**

Dataset collected constituted the demographic descriptors are described by frequency analysis as below.

**Fig.3:** shows the frequency analysis of the age. From the table it is clear that majority of respondents are age group 20 to 25 (225) followed by age group 25 to 30 (92), age group less than 20 (61) and age group greater than 30 (18).

**Fig.4:** shows the frequency analysis of the gender. From the table it is clear that the majority of the respondents are female (207) and male (189).

**Fig.5:** shows the frequency analysis of the educational designation. From the table it is clear that the majority of the students are undergraduates (234) followed by post graduates (158) and doctoral students.

**Fig.6:** shows the frequency analysis of the school board. From this table it is evident that the majority of respondents are of state board (296) and followed by central board (93) and others (7).
Table 1: ANOVA RESULTS

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Variable</th>
<th>F statistic</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V1</td>
<td>2.572448</td>
<td>0.037461</td>
</tr>
<tr>
<td>2</td>
<td>A1</td>
<td>3.051259</td>
<td>0.01699</td>
</tr>
<tr>
<td>3</td>
<td>R1</td>
<td>4.854752</td>
<td>0.000786</td>
</tr>
<tr>
<td>4</td>
<td>K1</td>
<td>5.062573</td>
<td>0.000549</td>
</tr>
<tr>
<td>5</td>
<td>V2</td>
<td>3.042365</td>
<td>0.017244</td>
</tr>
<tr>
<td>6</td>
<td>A2</td>
<td>0.536994</td>
<td>0.708639</td>
</tr>
<tr>
<td>7</td>
<td>R2</td>
<td>1.937822</td>
<td>0.103411</td>
</tr>
<tr>
<td>8</td>
<td>K2</td>
<td>2.605473</td>
<td>0.035493</td>
</tr>
</tbody>
</table>

From the above table it is inferred that V1, A1, R1, K1, V2 and K2 have an effect on style. A2 and R2 have no effect on style. We can interpret the style variable from the first part of the questionnaire as having an effect on the numerical descriptors.

Table 2: CHI SQUARE TEST RESULTS

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Variable</th>
<th>chi-square value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender</td>
<td>5.547826</td>
<td>0.235557</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>24.93714</td>
<td>0.015124</td>
</tr>
<tr>
<td>3</td>
<td>Place grownup in</td>
<td>30.4778</td>
<td>0.169416</td>
</tr>
<tr>
<td>4</td>
<td>Educational designation</td>
<td>5.668069</td>
<td>0.684357</td>
</tr>
<tr>
<td>5</td>
<td>school region</td>
<td>28.70998</td>
<td>0.231237</td>
</tr>
<tr>
<td>6</td>
<td>school Type</td>
<td>5.311445</td>
<td>0.723829</td>
</tr>
<tr>
<td>7</td>
<td>school board</td>
<td>11.96928</td>
<td>0.15258</td>
</tr>
<tr>
<td>8</td>
<td>college location</td>
<td>191.3169</td>
<td>0.001251</td>
</tr>
<tr>
<td>9</td>
<td>educational stream</td>
<td>64.91396</td>
<td>0.193906</td>
</tr>
</tbody>
</table>

From the above table it is evident demographic variables except age and college location have no effect on learning style. This proves that the initial assumption stating places and school background may have an effect on learning preference is wrong.

Table 3: MODEL RESULTS

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Method</th>
<th>accuracy</th>
<th>recall</th>
<th>precision</th>
<th>f1_score</th>
<th>Cross Val score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Logistic Regression</td>
<td>0.625</td>
<td>0.625</td>
<td>0.57357</td>
<td>0.57156</td>
<td>0.643924</td>
</tr>
<tr>
<td>1</td>
<td>Gaussian Naive bayes</td>
<td>0.55</td>
<td>0.55</td>
<td>0.51526</td>
<td>0.52799</td>
<td>0.563196</td>
</tr>
<tr>
<td>2</td>
<td>K Nearest Neighbour</td>
<td>0.6375</td>
<td>0.6375</td>
<td>0.50606</td>
<td>0.52290</td>
<td>0.507532</td>
</tr>
<tr>
<td>3</td>
<td>Support Vector Machines</td>
<td>0.65</td>
<td>0.65</td>
<td>0.67673</td>
<td>0.57272</td>
<td>0.621203</td>
</tr>
<tr>
<td>4</td>
<td>Decision tree</td>
<td>0.6</td>
<td>0.6</td>
<td>0.57686</td>
<td>0.58734</td>
<td>0.563038</td>
</tr>
<tr>
<td>5</td>
<td>Random Forest</td>
<td>0.7</td>
<td>0.7</td>
<td>0.74683</td>
<td>0.62761</td>
<td>0.648924</td>
</tr>
<tr>
<td>6</td>
<td>Gradient Boost</td>
<td>0.6875</td>
<td>0.6875</td>
<td>0.695</td>
<td>0.68169</td>
<td>0.669082</td>
</tr>
<tr>
<td>7</td>
<td>XG Boost</td>
<td>0.6875</td>
<td>0.6875</td>
<td>0.69300</td>
<td>0.67494</td>
<td>0.626329</td>
</tr>
<tr>
<td>8</td>
<td>Light GBM</td>
<td>0.6875</td>
<td>0.6875</td>
<td>0.73901</td>
<td>0.67402</td>
<td>0.661677</td>
</tr>
<tr>
<td>9</td>
<td>Extra trees Classifier</td>
<td>0.6375</td>
<td>0.6375</td>
<td>0.54290</td>
<td>0.54256</td>
<td>0.654051</td>
</tr>
<tr>
<td>10</td>
<td>Voting Classifier</td>
<td>0.7</td>
<td>0.7</td>
<td>0.73447</td>
<td>0.67559</td>
<td>0.686899</td>
</tr>
<tr>
<td>11</td>
<td>Neural Network classifier</td>
<td>0.62184</td>
<td>0.62184</td>
<td>0.55646</td>
<td>0.55111</td>
<td>0.67193</td>
</tr>
</tbody>
</table>

Model summary shows that voting classifier (with top 6 base models excluding Support vector machines) performs better than other models with accuracy (70%), recall (70%), precision (73%) and f1 score (67.5%). It is inferred that tree-based models perform better than the linear models and models give lower results than expected due to insufficient data points. Also, deep learning algorithms (Fully connected Neural networks with 2 hidden layers and optimizer stochastic gradient descent) performs better with more data.

SUGGESTIONS
As the online platform is being reformed as the norm in the educational sector, the learning styles classification helps personalize the learning experience. On that note the scope for the learning style classification is wide especially beyond the sensory modalities there are many behavioral traits that have to be included. Growth in the analytics sector gives away the readily available predictive algorithms. But algorithms that are specific to the student analytics or educational analytics may obtain better results regarding the classification of students based on their learning styles and traits. Processing the information based on the modalities applies for any human being, there is a gap in study of learning style in corporate training and development process that may be explored.
CONCLUSION
This study shows the ability of the machine learning algorithms to ascertain the relationships between the data. The learning styles are an important part of the student’s way of processing the information during the education. Voting classifiers may require the blend of computationally expensive algorithms to obtain better results. But better results trump that disadvantage. This study explored the relationship between demographic factors like school and place people grew up and learning styles. Results proved to contradict those factors. I conclude that with the growth of big data learning style classification, a blend of model algorithms or stacked algorithms like voting classifier can be used to adapt to a user application.

REFERENCES