Leaph

Using Data Mining and Machine Learning:

For academic performance improvement

Leaph.co

Abstract

The last 5 years have seen a total of 259 petabytes of data has been created and stored over the internet in various cloud systems, and this data continues to grow in volume as the world becomes even more connected than it was before. This is more data we could ever read in a life time and so to analyze this data a special field in the technology sector has been created; Data Science. The field of Data Science has irrevocably become one of mankind's most sort after tools when it comes to data analysis in content management systems, as such a great number of developers and scientists have taken a data mining approach to solving their problems.

Present day the volume of data stored in educational databases increasing rapidly. Accurately predicting a group students' performance has become more challenging due to the large, almost exponential increase of data in educational databases and systems. With these databases containing obscured information on improvement of students' performance. The performance in primary and secondary education is a turning point in the academics for all students, and this academic performance is influenced by many factors, therefore it is essential to develop predictive data mining models for students' performance so as to identify the difference between high learners and slow learning student. The single greatest challenge is to improve the quality of the educational processes so as to enhance student's performance. Therefore, it is crucial to set new strategies and plans for a better management to better develop effective teaching systems that benefit all categories of students, factoring in their learning types.

This will allow seamless integration of Aural, Logical, Verbal, Visual, Physical, Social and Solitary learners.

Current studies on existing prediction methods and systems is still too rudimentary to sufficiently identify the most effective and versatile methods for predicting the academic performance of students, and it goes without saying that there is also a great deficit of research on the factors a tha contribute to students' milestones and achievements in certain courses in the course of their academic lives. Therefore, an analytical review on predicting student performance by using data mining techniques is proposed to improve students' achievements in academia. The proposed research in this paper is intended to provide an overview and a current solution on the data mining techniques that have been used to predict students' performance. This paper also focuses on how the prediction algorithm can be used to identify the most important attributes in a students' data.

Our objective is to try to actually improve students' achievement and success significantly and effectively in an efficient method using educational data mining and machine learning techniques. We will look at the various ways we can achieve this and compare the individual results and models to find the most robust and efficient method we can use to solve this problem.

Introduction

Utilizing Data Mining to extract hidden predictive information from large databases, is a comprehensive new technology with untold potential to help educational institutions to pivot on the crucial information in their data servers. Data mining tools can predict future trends and behaviors', giving institutions the ability to make proactive, knowledge-driven decisions on how best to help students develop. By using automated, intended analyses offered by data mining to propel them beyond the analyses of past events provided by retroactive tools typical of decision support systems. Data mining tools can aid institutions in answering questions that traditionally were too tedious to resolve. Data science technology scours databases for hidden patterns, finding predictive information that administrative experts may miss because it lies outside their expectations. Data mining is a powerful tool for academic intervention. Using a tailored data mining system, an institution could, predict with 85+ percent accuracy which students will or will not graduate, what their' academic records show is their weakness, and how their' results determine regional exam scores. Institutions could use this information to focus on academic assistance on those students who are most at vulnerable or lacking.

With the rise of new Educational institutions being set up, institutions are becoming more competitive. To stay afloat, these institutions are focusing more on improving various aspects of the educational environment, and one important factor among them is the quality of learning. In order to provide quality education and brace for new challenges, all institutions need to know where their potentials lie both of which are explicitly seen and those that which are hidden. In today's educational institutions there is a substantial amount of knowledge is obscured from the institutions themselves. To get that competitive edge, these institutions must harness tools that help them identify their own hidden potentials and implement the best technique to bring it out. There has been an increase in in realm of educational data mining in recent years, as it has become a vital need for the academic institutions to improve the quality of education.

A Primer on Data Mining

So how does Data Science – Data Mining work? It's important to understand a few fundamental concepts. First, data mining relies on four essential methods:

- Classification
- Categorization
- Estimation
- Clustering
- Visualization

Classification: identifies associations and clusters and separates subjects under study. It is a technique in supervised learning, which generates a model to classify a data item conceding to a predefined class label. The intent of the classification scheme is to predict the future output based on available testing knowledge. Classification the prediction models could be used on the testing data on the students', and this data will narrate the official data of the students' performance in institutions. It can be used for acquiring a comprehensive analysis of student characteristics or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

Categorization: uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "stay."

Estimation: includes predictive functions and models to deal with continuous outcome variables, such as GPA, or test scores.

Visualization: uses interactive graphs to demonstrate mathematically induced rules and scores, it has far more depth and sophistication than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates. In this case however we will use visualization to show an overall result on students' performance using radar charts.

Toolkit

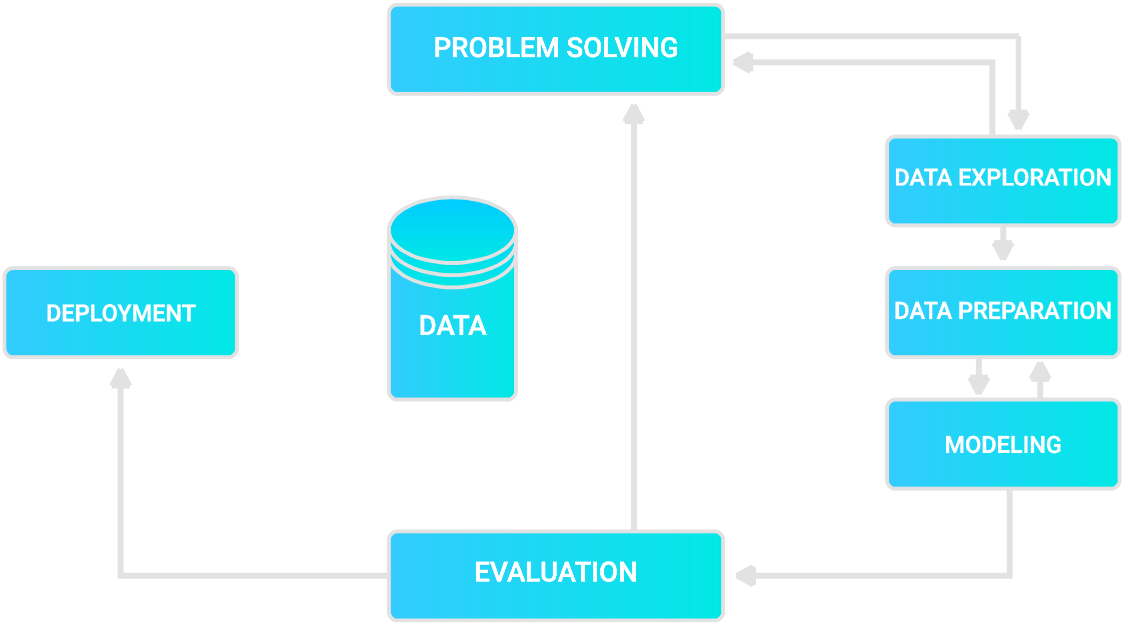
Data mining has a versatile toolkit to get the most out of its intended purpose. This shortens the time required to set up a tailored solution, among these are:

- Machine learning
- Neural networks
- Biological Engineering
- Induction algorithms
- Behavioral Psychology
- Computer science and heuristics
- Artificial intelligence
- Emulating human intelligence

Phases of Data mining

Data mining is an iterative process that typically involves the following phases:

- Problem definition
- Data exploration
- Data preparation
- Modeling
- Evaluation



To solve a data science problem, we start with the understanding of the problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. In Leaph™ our domain is academic data, the results of academic institutions over a vast period of years, strength of students per year and per department and the experts of that domain are Heads of Department of departments and principals of these institutions. In the data exploration phase, traditional data analysis tools primarily statistics, are used to explore the data. We then enter the data preparation phase, this is when preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The data is cleaned for missing data, isolated cases and tweaked multiple times in no prescribed order. The meaning of the data however remains unchanged.

Data Collection and Analysis Tools

Acquiring data requires various tools are needed to collect data for Leaph™, some for analyzing data, some for designing, implementation and some of these developing software tool these are:

- Excel
- MS Access
- J48 algorithm
- Naïve Bayesian Classifier
- WEKA data mining tool
- Tangara data mining tool
- Rapid Miner

Before we move on to describing how Leaph™ will use these tools and technologies here is brief analysis of why such a system should be used in schools:

- It'll help them identify weak or slower students.
- Gives a robust and diverse framework to aid in individual student development.
- Helps institutions create better syllabi that students can easily keep up with.
- Provides well-structured analytical information on both students and teachers.
- Exposes specific gaps or weaknesses per subject for each student.
- Fosters individual development for collective progress.
- Helps in predicting the result of students.

Leaph™

Description

Leaph™ is school management system designed to improve the student's performance using technologies mentioned above. It utilizes these technologies to map the students' records using a K-mean clustering algorithm and grouping datasets into cluster, and a prediction neural network system for performance prediction. Leaph™ provides an easy to use, intuitive and robust system educational institutions. It applies data mining techniques to identify whether students' learning experiences can be assessed based on their results but It is limited to the available data in online databases, excluding factors such as students' position in the collaborative group and Structure of the collaborative tasks is not considered.

Development

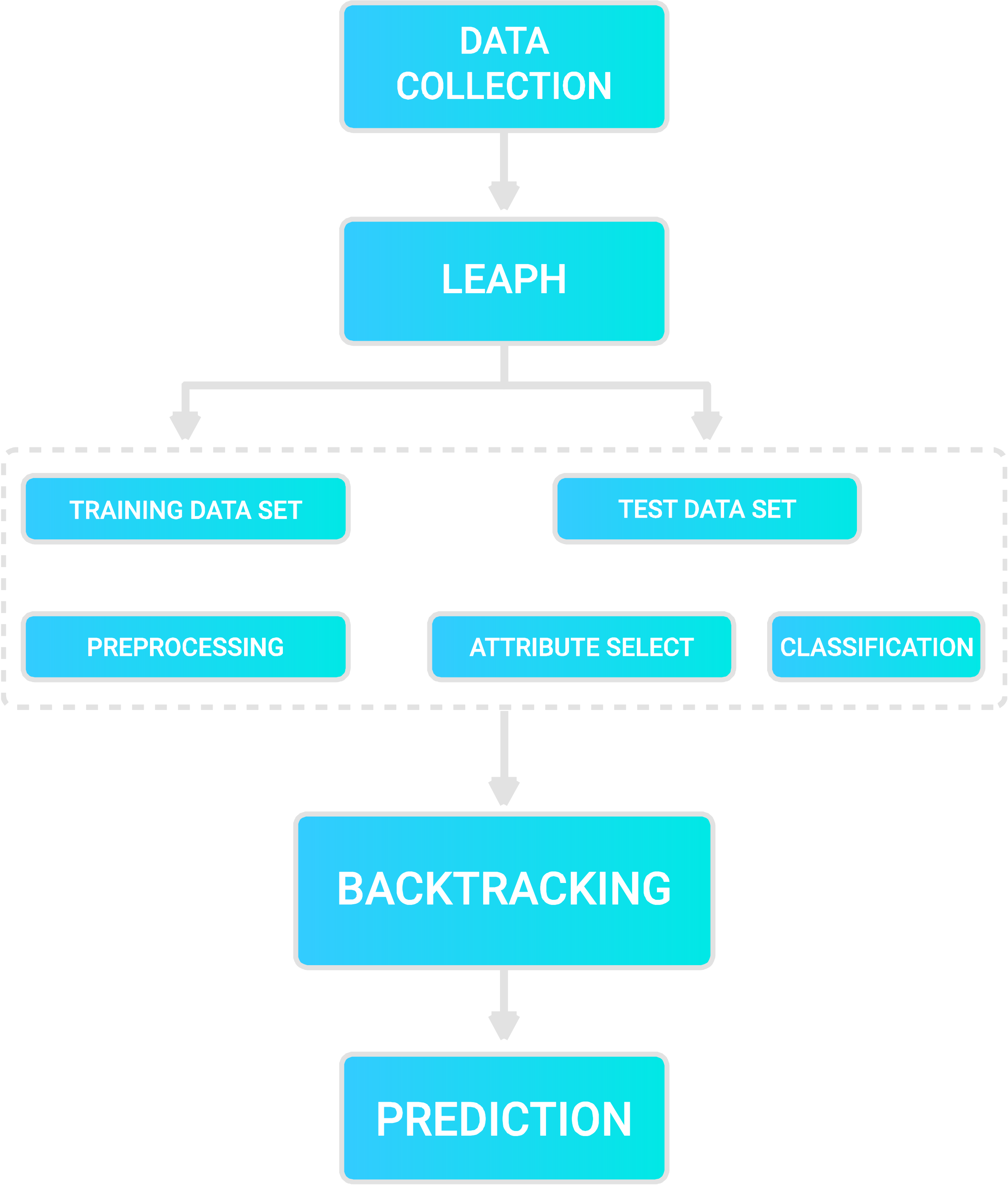
We are developing it in stages from the bottom up, starting with data visualization. Developed mostly in web technologies such as PHP, MySQL and Python Flask and visualizations in the Bokeh framework. Educational data mining has vast amount of data that has to be organized in a consistent manner. To organize, analyze and classify students details K-mean Clustering algorithm is been used based on academic records. Thereby forming three clusters based on students record.

- Low performance Student
- Average Student
- Smart Student

But it simply specifies the current scenarios whereas no future prediction is available and variables used for analysis are only based on demographic and academic records. Therefore, to enhance the existing system the proposed model is designed by collecting Students Personal and Academic data from the senior students of the institutions and thereby grouping the student's performance based on certain conditions:

- best
- good
- average
- poor

Student model is designed for the prediction of the outcome of the student based on the Leaph™ framework depicted below. This system provides an efficient analysis on student performance by data collection and result prediction.



Visualization

We have completed work on our visualization system yielding very good results as shown in the following snippets and prototypes.

Version 1

We use Python to render the visualizations. To begin load the libraries and data. Then use the `df.head()` to view the structure of the dataset. We will be using Matplotlib, Pandas, Seaborn, Numpy, and using test data generated from here and random values below 600 in Excel for students test scores.

```
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np
df=pd.read_csv("C://Users/Fred/Desktop/Projects/Leaph/Students.csv")
df.head()
```

#	Name	Total	Logic	Abstract Thought	Creativity	Memory Retention	Strategy	Communication	Grade
1	Annie Moses	318	45	49	49	65	65	45	C
2	Fletcher Kirk	405	60	62	63	80	80	60	B
3	Miguel Velasquez	525	80	82	83	100	100	80	A
4	Laylah Greer	625	80	100	123	122	120	80	A
5	Desiree Thornton	309	39	52	43	60	50	65	C

The Radar Chart

As shown above the data contains 7 variables for each student. So after we know our dataset, it's time to draw the radar chart, I'm choosing this one because it is the most efficient way to visualize this data. We want to show the **'Logic', 'Abstract Thought', 'Creativity', 'Memory Retention', 'Strategy', 'Communication'** as 6 different axes on our radar chart, so just take them out and set as a np.array. Here we use Student No.386 "Caylee Weaver" as an example to illustrate the chart.

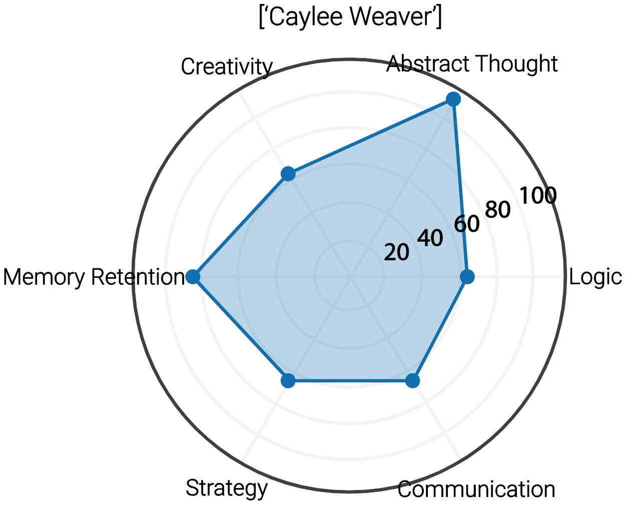
```
labels=np.array(['Logic', 'Abstract Thought', 'Creativity', 'Memory Retention', 'Strategy', 'Commun
ication'])

stats=df.loc[386,labels].values
```

Set the angle of polar axis. And here we need to use the np.concatenate to draw a closed plot in radar chart.

```
angles=np.linspace(0, 2*np.pi, len(labels), endpoint=False)

# close the plot

stats=np.concatenate((stats,[stats[0]]))

angles=np.concatenate((angles,[angles[0]]))
```

Here we should use the fig.add_subplot rather than the sns.plt.subplots().(notice the "s"). Because the subplots doesn't contain the argument "polar". We can only set the polar axis by subplot.
Then draw the plot as the frame and fill in the surrounded area by fill(). At the end set the label of axis and the title then everything done. Keep in mind this project book is just to allow me to document and quickly prototype parts of the system as proof of concept.

```
fig=plt.figure()

ax = fig.add_subplot(111, polar=True)

ax.plot(angles, stats, 'o-', linewidth=1)

ax.fill(angles, stats, alpha=0.15)

ax.set_thetagrids(angles * 180/np.pi, labels)

ax.set_title([df.loc[386,"Name"]])

ax.grid(True)
```

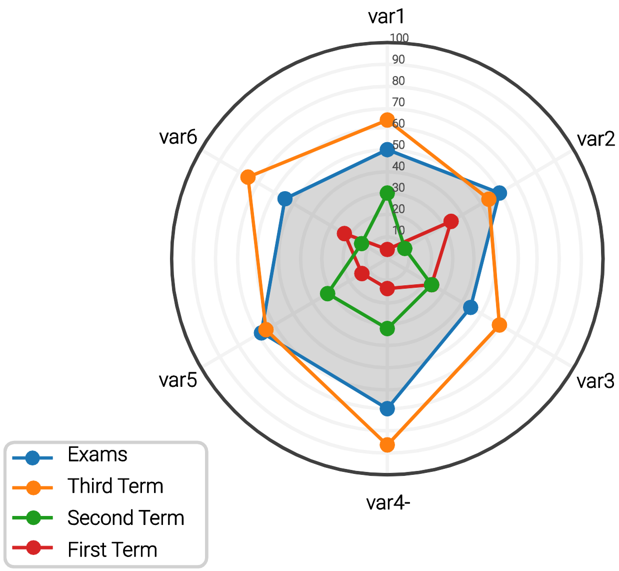


Improving the Visualization

We believe learning is about self-development and mastery not test scores or grades. Students should be scored against their previous development not their peers. To effectively integrate this into the system we added another dimension to this chart, to show a student's improvement plotted against their previous results. This will help parents, teachers and students to visualize how they are performing and exactly where they are falling short.

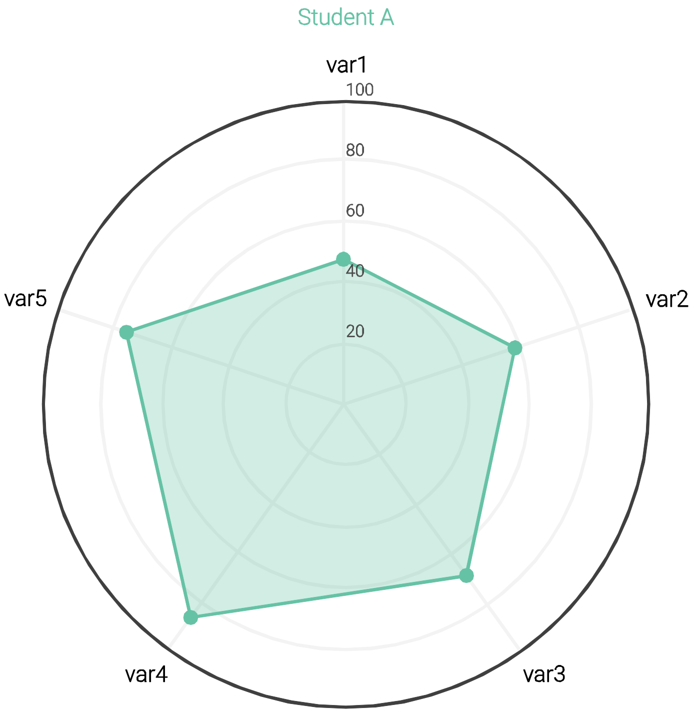
```
# Libraries
import matplotlib.pyplot as plt
import pandas as pd
from math import pi

# Setting the data, For now it's all fake but this data will load from sql and or json + .csv files on a server or through an API.
df = pd.DataFrame({
'group': ['A','B','C','D'],
'var1': [50, 65, 30, 4],
'var2': [60, 55, 9, 34],
'var3': [45, 60, 23, 24],
'var4': [70, 87, 33, 14],
'var5': [67, 65, 32, 14],
'var6': [55, 75, 14, 23]
})



# ------- PART 1: Creating background

# number of variable
categories=list(df)[1:]
N = len(categories)

# What will be the angle of each axis in the plot? (we divide the plot / number of variable)
angles = [n / float(N) * 2 * pi for n in range(N)]
angles += angles[:1]

# Initialise the spider plot
ax = plt.subplot(111, polar=True)

# If you want the first axis to be on top:
ax.set_theta_offset(pi / 2)
ax.set_theta_direction(-1)

# Draw one axe per variable + add labels labels yet
plt.xticks(angles[:-1], categories)
```

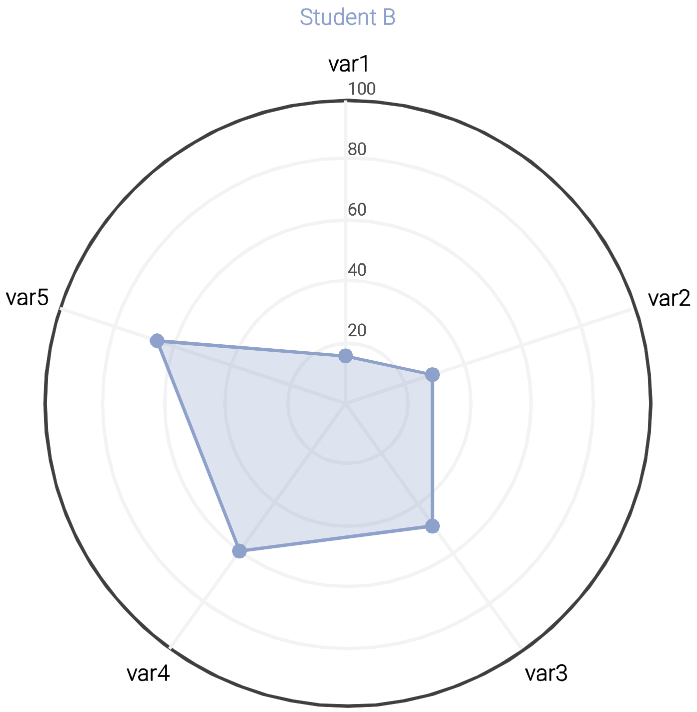
```
# Draw ylabels
ax.set_rlabel_position(0)
plt.yticks([10,20,30,40,50,60,70,80,90,100], ["10","20","30","40","50","60","70","80","90","100"],
color="grey", size=7)
plt.ylim(0,100)



# ------- PART 2: Add plots

# Plot each individual = each line of the data
# I don't do a loop, because plotting more than 3 groups makes the chart unreadable

# Ind1
values=df.loc[0].drop('group').values.flatten().tolist()
values += values[:1]
ax.plot(angles, values,'o-', linewidth=1, linestyle='solid', label="Exams")
ax.fill(angles, values, 'b', alpha=0.1)

# Ind2
values=df.loc[1].drop('group').values.flatten().tolist()
values += values[:1]
ax.plot(angles, values,'o-', linewidth=1, linestyle='solid', label="Third Term")
ax.fill(angles, values, 'r', alpha=0.05)

# Ind3
values=df.loc[2].drop('group').values.flatten().tolist()
values += values[:1]
ax.plot(angles, values,'o-', linewidth=1, linestyle='solid', label="Second Term")
ax.fill(angles, values, 'c', alpha=0.05)

# Ind4
values=df.loc[3].drop('group').values.flatten().tolist()
values += values[:1]
ax.plot(angles, values,'o-', linewidth=1, linestyle='solid', label="First Term")
ax.fill(angles, values, 'r', alpha=0.05)

# Add legend
plt.legend(loc='upper right', bbox_to_anchor=(0.1, 0.1))
```

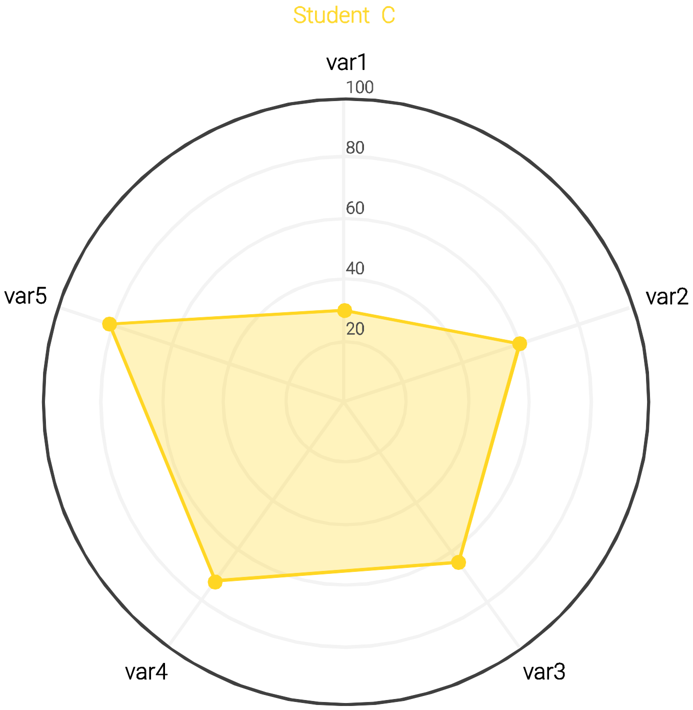


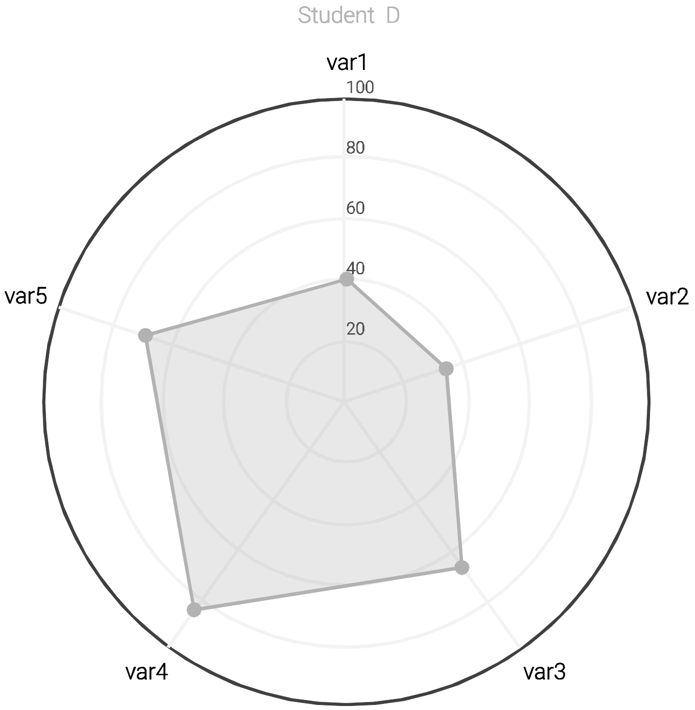
Using our visualizations; teachers can easily read and understand the complex data produced by the system. They can also use this data to see if their methods are working or to quickly see which students are performing the worst and take note.

```
#### Libraries
import matplotlib.pyplot as plt
import pandas as pd
from math import pi

# Set data
df = pd.DataFrame({
'student': ['A','B','C','D'],
'var1': [48, 15, 30, 40],
'var2': [59, 30, 60, 34],
'var3': [68, 49, 63, 64],
'var4': [87, 61, 73, 85],
'var5': [75, 65, 82, 70]
})

# ------- PART 1: Define a function that do a plot for one line of the dataset!

def make_spider( row, title, color):

    # number of variable
    categories=list(df)[1:]
    N = len(categories)

    # What will be the angle of each axis in the plot? (we divide the plot / number of variable)
    angles = [n / float(N) * 2 * pi for n in range(N)]
    angles += angles[:1]

    # Initializing the spider plot
    ax = plt.subplot(2,2,row+1, polar=True, )

    # If you want the first axis to be on top:
    ax.set_theta_offset(pi / 2)
    ax.set_theta_direction(-1)

    # Drawing one axe per variable + adding labels labels yet
    plt.xticks(angles[:-1], categories, color='grey', size=8)

    # Drawing ylabels
    ax.set_rlabel_position(0)
    plt.yticks([20,40,60,80,100], ["20","40","60","80","100"], color="grey", size=7)
    plt.ylim(0,100)

    # Ind1
    values=df.loc[row].drop('student').values.flatten().tolist()
    values += values[:1]
    ax.plot(angles, values,'o-', color=color, linewidth=1, linestyle='solid')
    ax.fill(angles, values, color=color, alpha=0.3)

    # Adding a title
    plt.title(title, size=11, color=color, y=1.1)

# ------- PART 2: Apply to all individuals
# initializing the figure
my_dpi=96
plt.figure(figsize=(1000/my_dpi, 1000/my_dpi), dpi=my_dpi)

# Creating a color palette:
my_palette = plt.cm.get_cmap("Set2", len(df.index))

# Loop to plot
for row in range(0, len(df.index)):
    make_spider( row=row, title='student '+df['student'][row], color=my_palette(row))
```









We've completed our data visualization for score prediction based on test scores and quizzes. We plan on utilizing a natural language processing algorithm to analyze what the teacher is thinking to improve about the student or if the student is improving at all regardless of the data. The prediction will be done using a prediction algorithm that works by finding patterns in the students test scores, then adjusting for errors. The real problem is test data. Most good schools have predictions for their O'level and A'level students based on previous scores. These predictions however shouldn't be looked at as a destiny but rather as encouragement. Neither should these predictions be used as comfort.

Further improvements of this system can be the "parallel marking," scanned tests can be analyzed and digitized for this program this will allow it to map exactly where a student might be going wrong or what they may have missed. This will allow for targeted assignments and notes to aid students and help them catch up with the rest of the syllabus. These algorithms aren't too great a challenge to implement, it's acquiring the data and training these algorithms that is tedious and resource intensive.

At this point the system is already 25% complete.

The next stage will be designing and implementing the data analysis system and training our machine learning models based on this data. This part of development has been throttled but lack of infrastructure. To train our models and test the system we need access to a computational power resource or cloud framework, which is currently out of our budget.

Conclusion

Current educational systems do not integrate any data mining or machine learning tools, to predict students fail or pass percentage based on the performance. The system doesn’t deal with privately tutored students, nor does it take into consideration that different students have different learning styles and have a certain optimum learning environment to study. There isn’t an efficient method to caution the administrative departments about the students’ deficiencies and lack of attendance. It doesn’t identify the weak student and inform the teacher. Another common problem in larger colleges and universities, some students may feel lost in the crowd or completely isolated in the learning environment. Whether they’re struggling to find help with coursework, or having difficulty choosing (or getting into) the courses they need, many students are daunted by the task of working through the collegiate bureaucracy. Leaph™ will aid in identifying the weak students, to caution teachers so they can provide academic help for them. It also helps the teacher to act before a student drops or plan for recourse allocation with confidence gained from knowing how many students are likely to pass or fail. Leaph™ will also show data graphically according to the need or organization which helps them understand it and to make important decisions considering their Students. For future work we also use clustering and neural networks. With the help of clustering we can see the domain and interest of students in particular field, and with neural networks we can theoretically reach an optimum 95% accuracy in prediction and anomaly detection among student results.

Copyright and license