JULY 2021

Data Science Final Project

Computer Science & Entrepreneuership
Semester 2

PREPARED BY

Yuval Bauberg
Jade Derhy
Tohar Nissan
Hadas David
Matanel Man

Introduction to Data Science

Phase 1:

This phase is divided into two parts, which is very similar to our process as a group.

First Attempt :

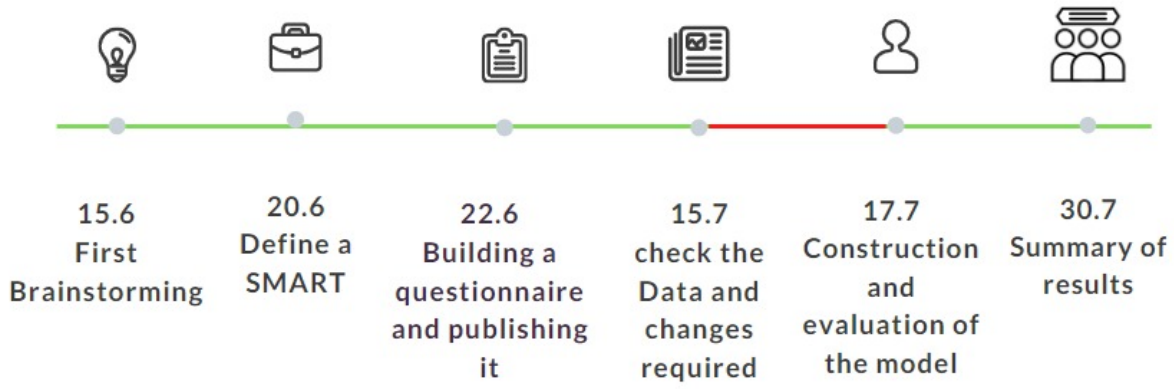
The purpose of our first attempt was to investigate the relationship between double major entrepreneurship students with certain characteristics and the second major they choose - computer science, economics, or business administration.

In order to collect our data, we have created a Google Form questionnaire based on two main topics:

1. Previous background: Army service, volunteer experience, SAT grades, etc.
2. Characteristic traits: after consulting with the Vice dean, dr. Yossi Maaravi, we chose 5 main characteristics based on the article: “Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners’ personality traits, business creation, and success” by Andreas Rauch & Michael Frese.

We managed to receive 203 results after a month of hard work and consistency.

Our goal was to find a connection between those answers to the choice of studying the second major (CS, Business management & Economics).
We tested, for example, if students with a business already chose to study Business Administration or if students in the 8200 unit chose to study Computer Science.



Phase 2:

In order to have an arranged data, we made all of our questions mandatory and for the numerical questions, we divided them by range.
In this way, we avoid blank space in our data, as seen below.

```
df.isna().sum()
```

```
ID                          0
field_of_study              0
Age                         0
Sex                         0
SAT_grade                   0
Scouts_instructors          0
Army_field                  0
Officer_carrier             0
_study                      0
Business_during_study       0
SelfEsteem                  0
Proactive                   0
Creativity                  0
Social                      0
```

The columns we have used are:

- field_of_study
- Age
- Sex
- SAT_grade
- Scout_instructor (Binary answer)
- Army_filed
- Officer_carrier (Binary answer)
- Business_before_study (Binary answer)
- Business_during_study (Binary answer)
- Self-esteem (from 1-5)
- Proactive (from 1-5)
- Creative (from 1-5)
- Social (from 1-5)

In each case, students will be presented by a vector of 13 values representing the characteristics, with each value classified by 0 (if it is No) or 1 (if it is Yes), if the values are multiple choices, we number them from 0 to the number of answers.

Phase 3:

The model we chose to focus on is K - means algorithm.

The purpose of this study is to develop a model for classifying students in the IDC entrepreneurship major into clusters with low internal variability and high external variability.

If we were able to discover these similarities and differences, we could identify whether or not students who choose entrepreneurship majors tend to share the same characteristics.

The steps are as follows:

1. We define the initial four centroids randomly
2. We calculate the distance between each point and each centroid.
3. For each data point, we calculate the shortest distance (from the distances we calculated in the previous step) and use this value to calculate the centroids of each cluster.

→ The goal is to minimize the total square distance between students and their centroids.

→ Decision variables - centroid values (13 values * 4 centroids = 52 decision variables in total)

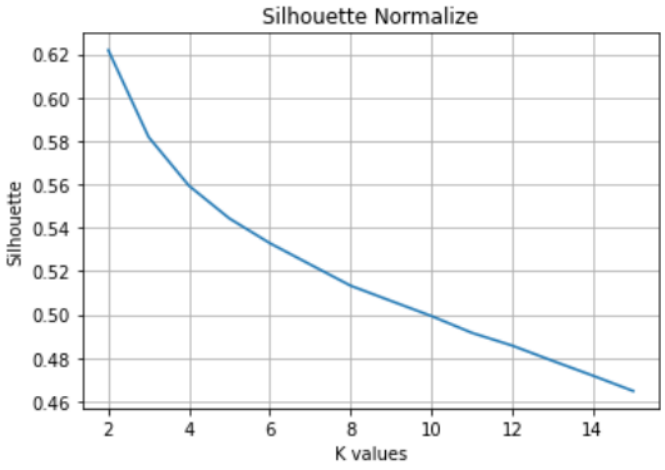
The results we got :

Surprisingly we found out that all of our 203 subjects belong to the same cluster.

As we can see in the graph we do not get the elbow we were hoping for.

As a result of this attempt, we can see that students who study Entrepreneurship have many things in common.

We then decided to do our research on a different question.



Attempt 2 :

Phase 1:

In light of our previous results and the fact that we collected and preprocessed the data ourselves, we decided to keep the data and the target the same.
After brainstorming, we came up with an interesting research question that we will be able to evaluate using the current data.
We decided to focus on our variable “Army_field” and to do our research on a new question.
During this part of our study, the students were asked about their service units, whether or not they served in an intelligence unit.

Research background:
Living in a country like Israel which has a mandatory army and the famous title of the “Start-up nation” Intelligence units such as 8200 and 8153 are “the place to serve”.
Startups and high-tech companies compete for those special graduates.
The research was inspired by the Calcalist[1] supplement's mapping who conducted a survey showing that soldiers and officers from Unit 81 in 2003-2010 have established a countless number of start-ups in the decade since their release: Approximately 100 graduates of the unit have established at least 50 companies.
In addition, The "Gotfriends" research[2] results show an increase in demand for 8200 alumni. If in 2017 the demand was 16% more for 8200 alumni in comparison to workers without military experience, then in 2018 the demand has already jumped to 24% and it continues to grow.
Alumni from these units earn about 20% above the average salary in the market for identical positions.
Another interesting statistic that emerges from the report of the placement company “see.V”[3] Graduates of technology units receive inquiries from about 20 different companies.

Our research question: “If he/she served in a special intelligence/technology unit, what are his/her chances of working at a startup or a high-tech company, or be an entrepreneur based on his/her professional background and character".

The success of our research will be determined by the objective variable "Army_field" (‘0’ if you served in a military intelligence unit). We are hoping to find out which of our students is more likely to work for a startup or high-tech company.

---

[1] https://newmedia.calcalist.co.il/magazine-07-01-21/m01.html

[2] https://www.gotfriends.co.il/%D7%91%D7%9C%D7%95%D7%92%D7%99%D7%9D/%D7%91%D7%95%D7%92%D7%A8%D7%99-8200-5-%D7%A2%D7%95%D7%91%D7%93%D7%95%D7%AA-%D7%A9%D7%90%D7%AA%D7%9D-%D7%97%D7%99%D7%99%D7%91%D7%99%D7%9D-%D7%9C%D7%93%D7%A2%D7%AA-%D7%95%D7%92%D7%9D-%D7%9C%D7%94%D7%AA%D7%92%D7%90%D7%95%D7%AA/

[3] https://www.geektime.co.il/see-v-tech-units/

In order to obtain the most accurate and realistic results, we used some articles, the data we collected, and the tools we learned in class during the semester.

On this project we chose to work together as a team, as we found the process very interesting and educational.

Having just two weeks to prepare, we were a little rushed, but with the help and advice of our TA, Mr. Efi Pecany, we knew it was possible.

Phase 2:

Based on our research and the connection we found between serving as a member of an intelligence or technology unit and working at high-tech companies or being an entrepreneur, we have decided to focus our efforts on adjusting our data for the new question.

A taste of our data:

	ID	field_of_study	Age	Sex	SAT_grade	Scouts_instructors	Army_field	Officer_carrier	_study	Business_during_study	SelfEsteem	Proactive	Creativity
0	1	0	0	0	0	1	0	1	1	1	2	2	2
1	2	0	1	0	6	1	0	1	0	0	3	5	3
2	3	2	2	0	5	0	0	0	0	1	5	4	5
3	4	2	1	0	8	0	2	0	0	0	4	3	5
4	5	2	1	0	5	1	2	0	1	1	5	5	5

Details of the data statistics:
This statistics was used in an effort to find some insights in our data.

	ID	field_of_study	Age	Sex	SAT_grade	Scouts_instructors	Army_field	Officer_carrier	_study	Business_during_study
count	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000	203.000000
mean	102.000000	1.004926	1.334975	0.413793	6.152709	0.492611	0.684729	0.339901	0.157635	0.221675
std	58.745213	0.858836	0.618292	0.493730	1.832417	0.501181	0.465772	0.474846	0.365300	0.416400
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	51.500000	0.000000	1.000000	0.000000	5.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	102.000000	1.000000	1.000000	0.000000	6.000000	0.000000	1.000000	0.000000	0.000000	0.000000
75%	152.500000	2.000000	2.000000	1.000000	8.000000	1.000000	1.000000	1.000000	0.000000	0.000000
max	203.000000	2.000000	3.000000	1.000000	8.000000	1.000000	1.000000	1.000000	1.000000	1.000000

We used the collected data in addition to the information in phase 2.
In this attempt, only the "Army_field" column was preprocessed.
We changed the values to binary values, ‘0’ if you served in an intelligence & technology unit and ‘1’ etc.

Phase 3:

The model we chose to focus on is decision tree algorithm:

Decision tree is a Machine Learning algorithm for solving the classification problem, in which the solution is represented by a decision tree. It offers a step-by-step process for making predictions and has a number of potential advantages over linear regression:

1. They are easily understood by non-experts and are compatible with most people's view of special issues.
2. There is no requirement that the relationship between destinations with similar characteristics be linear.
3. Searching for the best prediction properties is done automatically by the tree.
4. The decision tree is less sensitive to abnormal observations than regression.

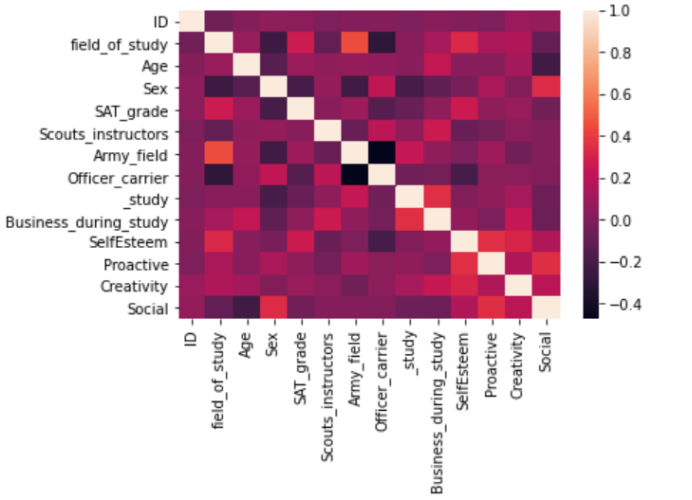
The decision tree algorithm selects the root property with the greatest expected information gain, while at the following nodes, the algorithm selects the property (not yet selected) with the greatest expected information gain. Whenever a threshold level is present, the algorithm determines the optimal threshold level for each property (i.e., the threshold level which maximizes the expected information gain for that property) and begins the calculation.
In categorical properties, the information obtained is usually based on the label of the property; in our project the label is "Army_field".
Using the "maximize profit from information" criteria defined above, the decision tree algorithm determines the optimal root node of the tree. Similarly, the decision tree algorithm continues to do so for the subsequent nodes. Among the leaf nodes of the decision tree are the probabilities that each category is the correct one, since it is used for classification.

Assumptions:

It is well known that there is a large percentage of high-tech and start-up workers in Israel that used to serve in intelligence units. Therefore we could assume that individuals that served in intelligence units and also chose to study a technological degree that combines entrepreneurship, are more likely to later on be entrepreneurs, establish their own start-up company or work in an existing high-tech company than others.
Therefore, we chose to use a Decision Tree as a classifying algorithm that could provide insight by what criteria it is easiest to classify someone as a future startup manager or employee.

We used seaborn heatmap in order to predict the correlation between the different features. We see that the features are correlated with each other.

```
sns.heatmap(df.corr());
```



The visual representation of our model is:

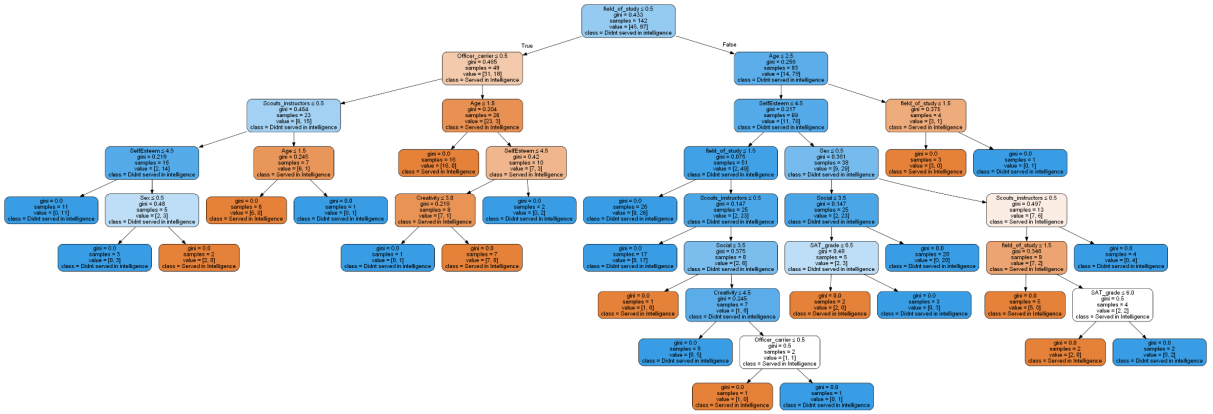
```
cols = list(df.columns)
cols.remove("Army_field")

X = df[cols].copy()
y = df["Army_field"].copy()
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

```
tree_clf = DecisionTreeClassifier(random_state=42)
```

```
tree_clf = tree_clf.fit(X_train, y_train)
```

```
plt.figure(figsize=(25,15))
plot_tree(tree_clf,filled=True,rounded=True,class_names = ["no", "Yes"], feature_names=X.columns);
```

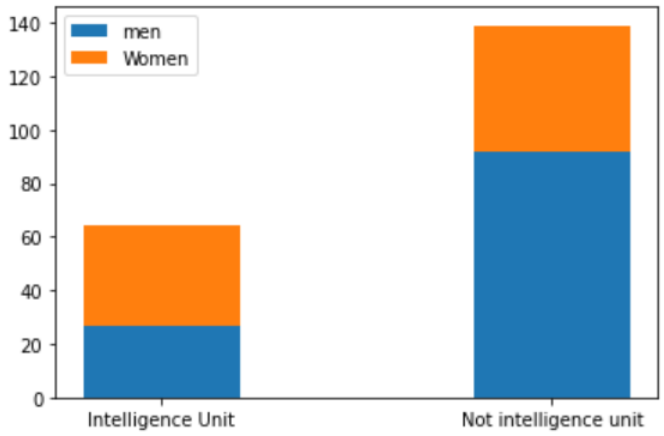


Conclusions from the decision tree we got:

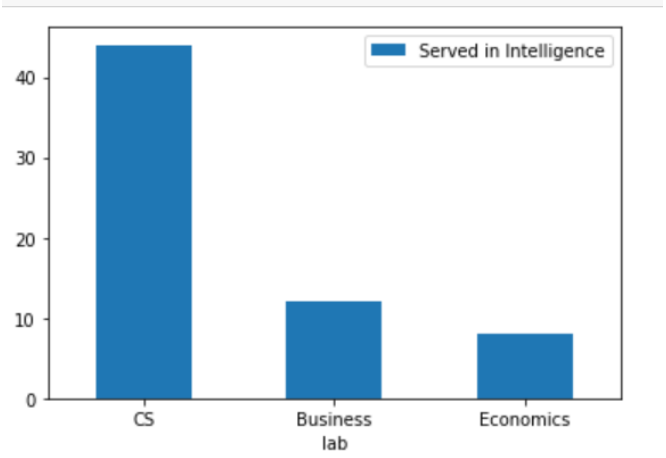
The conclusions were divided into two main parts (as we wished). The first group of conclusions represents students who served in an intelligence/technology unit, and the second group represents students who did not serve in an intelligence/technology unit.

Some interesting insights from the data:

As we can see in the plot below, more womens served in an intelligence unit.



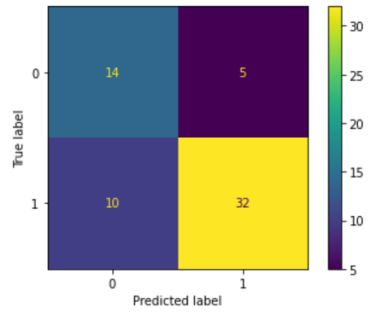
Most of the students who graduated from an intelligence unit chose to study Computer Science and not Business and Economics.
We can learn from this plot and the articles we mention above that the students who served in Intelligence units and study Computer Sciences are more likely to find a job in an High-Tech company or to fund their own startup company.



Our model evaluation:

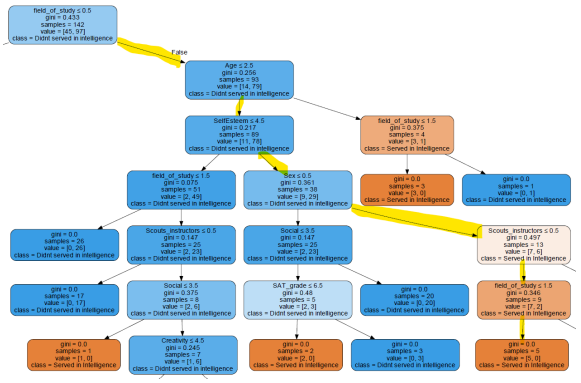
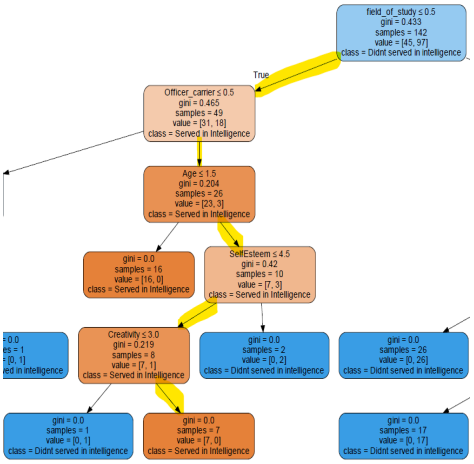
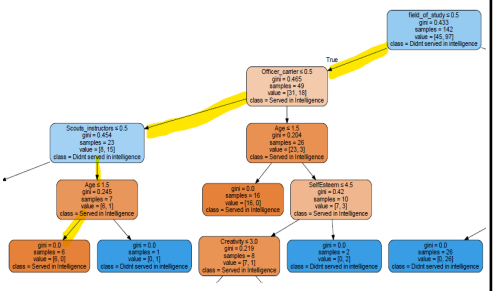
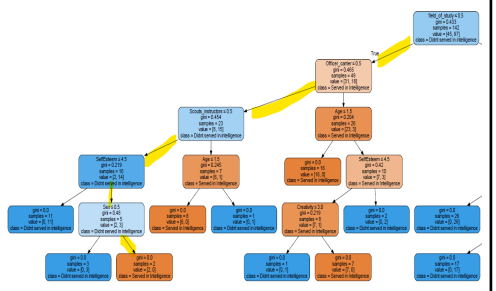
```
plot_confusion_matrix(tree_clf, X_test, y_test, display_labels = ["0", "1"])
```

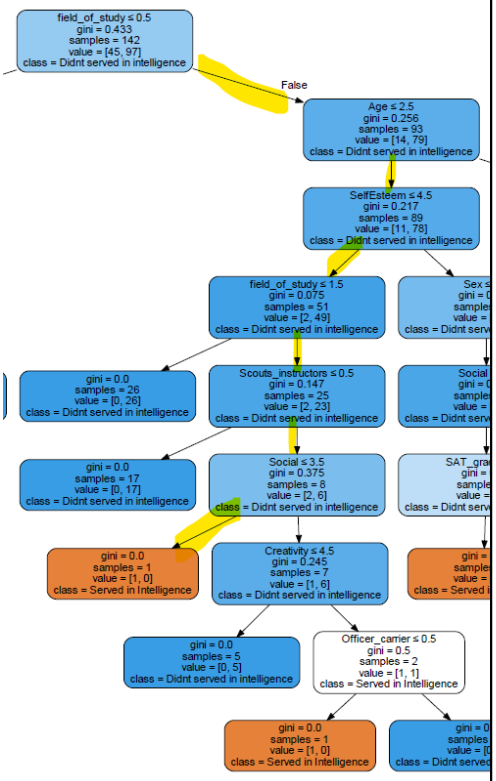
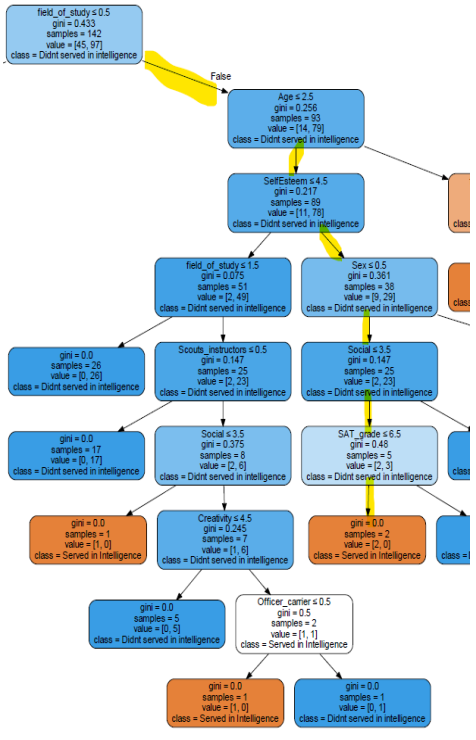
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1ada1a91a90>
```



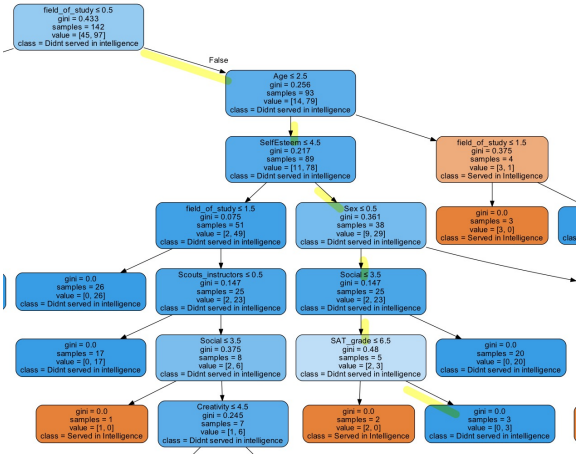
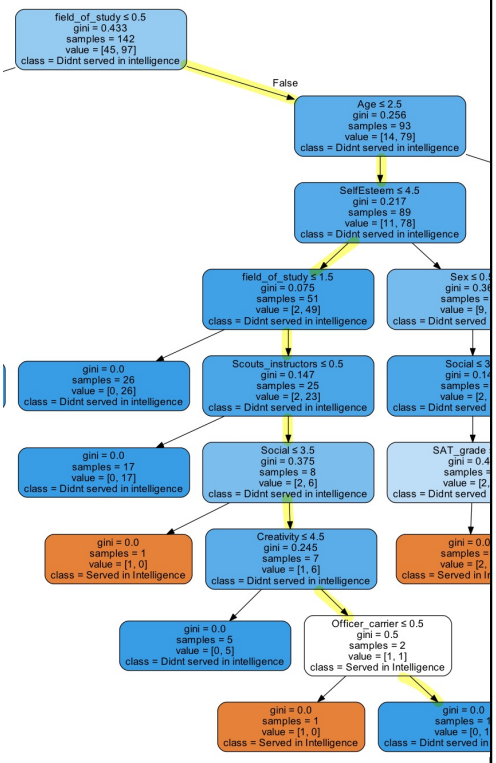
In the confusion matrix, we see that 14+5=19 people did not serve in an intelligence/technology unit, 14 were correctly classified.
In addition, 10+32=42 students served in an intelligence/technology unit, 32 were correctly classified.

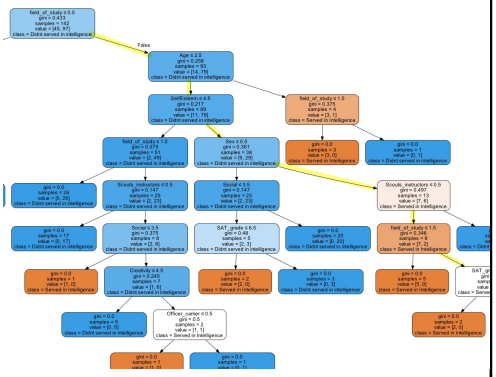
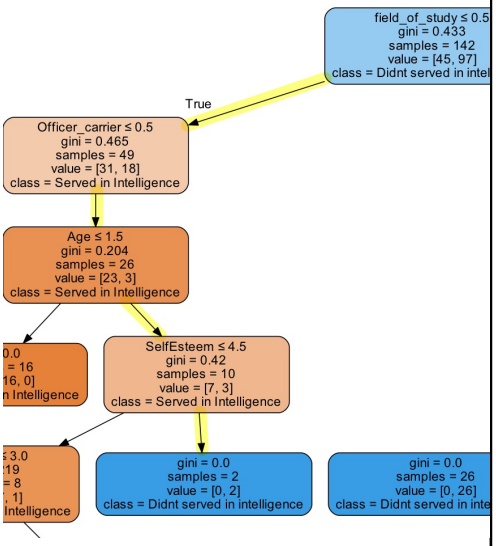
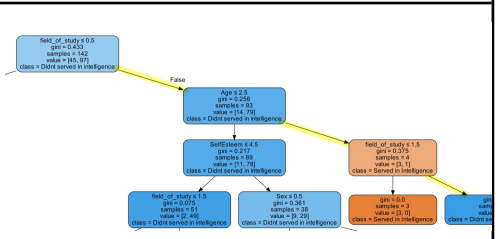
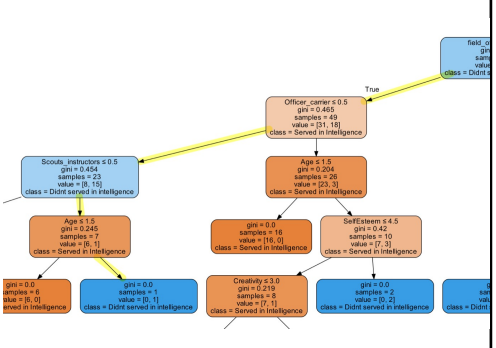
Students who served in an intelligence/technology unit are more likely to be:

Features	Path in the tree
Female economics and business management students under the age of 28 with a self-esteem of at least 4.5, who are not scout instructors.	
Students who were officers in the army and are older than 25, that also have lower self esteem from 4.5, and creativity higher than 3, served in an intelligence technology unit.	
CS and business management that were scouts instructors, and are less than 25 years old.	
Female business management and CS students that we not scouts instructors, and that have self-esteem greater than 4.5.	

Economics students that are younger than 28, were scouts instructors and have a social skills lower than 3.5 served in an intelligence unit.	
Business and economic students older than 28 with a social skills score of less than 3.5 and a SAT score less than 750.	

Students who did not serve in an intelligence/technology unit are more likely to be:

Features	Path in the tree
Male Business management and economic students under the age of 28 with self-esteem higher than 4.5, social skills under 3.5, and SAT score under 750.	
Business management students under the age of 28, with self-esteem lower than 4.5 who were not scouts instructors with social skills greater than 3.5, and creativity greater than 4.5, and who were officers in the army.	

Female economic above the age of 28, with SAT score greater than 700, who were scouts instructors, with self-esteem higher than 4.5.	
CS students above the age of 26 with self esteem greater than 4.5 that were not officers in the army.	
Business management students above the age of 26.	
CS students above the age of 26 who were not scouts instructors.	

Conclusion

We collected the data from all years of entrepreneurship students of the “Adelson” school of Entrepreneurship.
We managed to receive 203 results.
We started our process to try to use the K-means algorithm in order to classify the students into clusters.
The results of the algorithm were not good enough. We tried to divide the results into 4 groups but we received only one cluster.
The main reason for that is that we actually really took one group of similar people, which is probably the reason they all got accepted into the Adelson school for Entrepreneurship.
After consulting with our TA, we changed our research question and decided to focus on the chances of an entrepreneurship student to get a job at a start-up/high-tech/be an entrepreneur.
We then used the decision tree algorithm in order to learn and classify our data.
In phase 3 we presented some interesting plots and insights on students who served in an intelligence/technology unit and students who did not.
We believe students who served in 8200 or 8153 etc have a greater chance of working in the industry, but also students who did not serve in those units who are willing to work hard and have the right skills and background will be able to get a job at a high-tech company or start-up.