Τεχνική Νοημοσύνη - Εργασιά 2

Ονοματεπώνυμο	A.M
Ηλίας-Θεοφάνης Γραββάνης	3200248
Μάριος Γόρδιος	3180036

Στην εργασία υλοποιήθηκαν οι αλγόριθμοι Naive Bayes και ο ID3

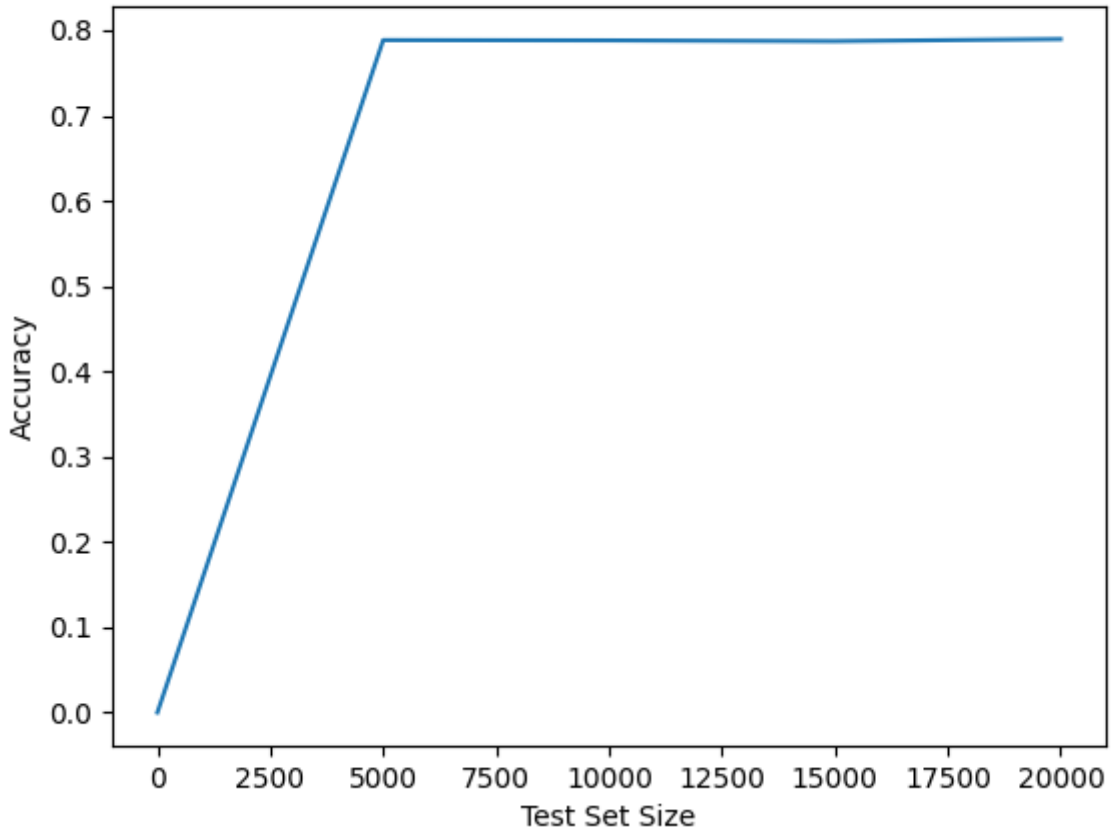
Μέρος Α

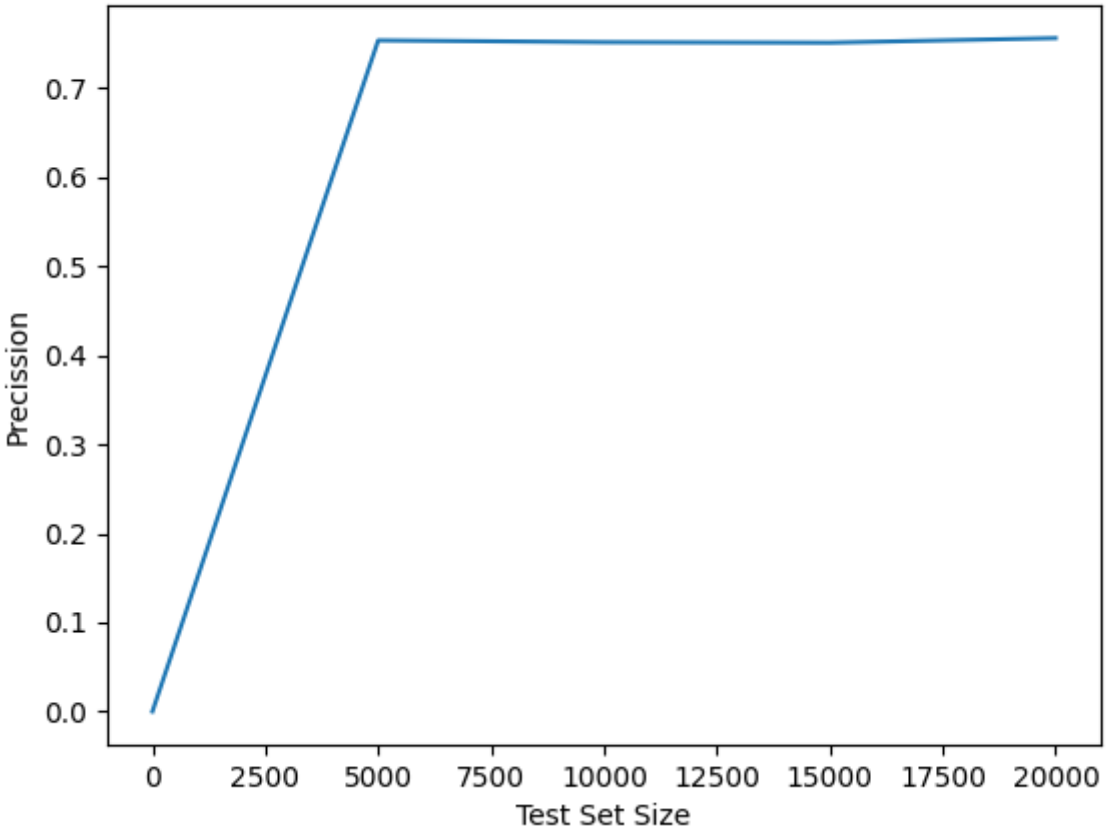
Αλγόριθμος Naive Bayes

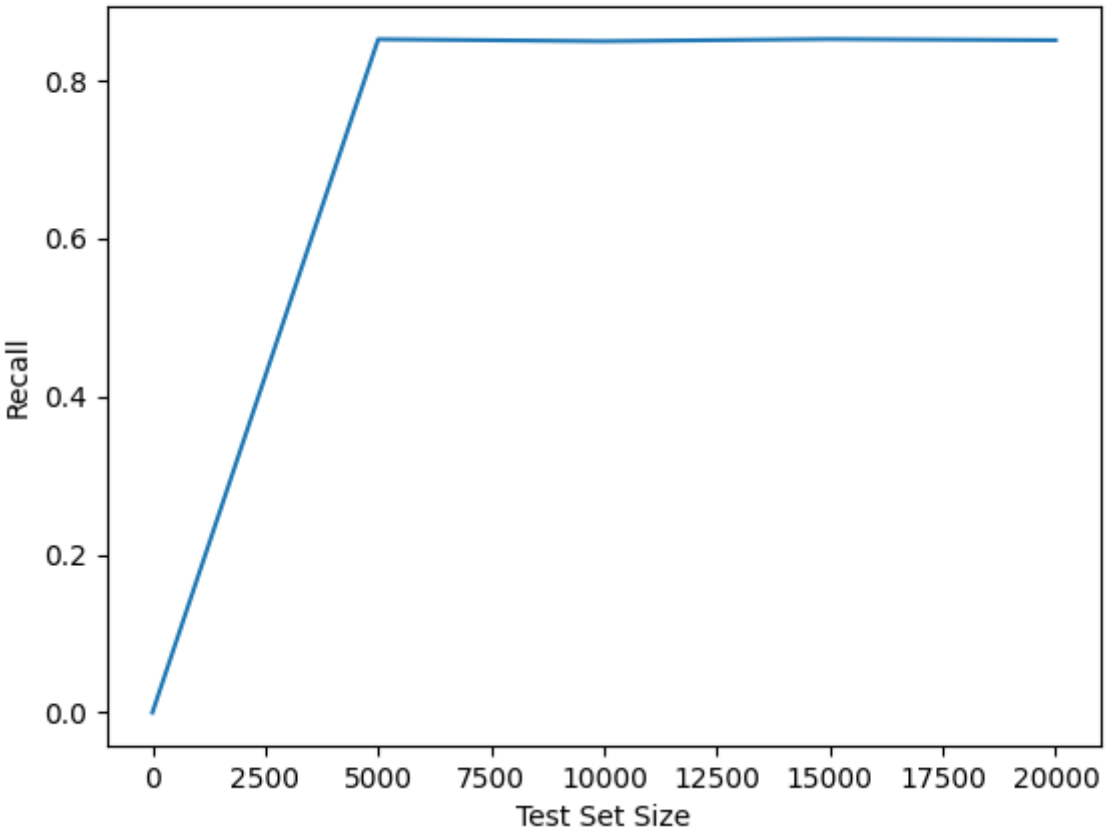
Στην υλοποίηση μας του αλγορίθμου Naive Bayes αρχικά παίρνουμε τις κριτικές των δεδομένων εκπαίδευσης για να δημιουργήσουμε ένα λεξικό, στο οποίο η κάθε λέξη έχει μοναδικό αναγνωριστικό (id) και συχνότητα, βάσει στην οποία θα υπολογιστούν οι πιθανότητες. Στη συνέχεια διαβάζουμε τα δοκιμαστικά δεδομένα (test data) και παίρνουμε ετικέτες για κάθε κριτική, χρησιμοποιώντας το λεξικό από τα δεδομένα εκπαίδευσης (training data) και τις πιθανότητες τους, χρησιμοποιώντας και τους προαναφερθέντες πίνακες συχνοτήτων. Για να υπολογίσουμε όλες τις πιθανότητες χρησιμοποιούμε την συνάρτηση log_sum η οποία διασφαλίζει ότι δεν θα προκύψει αριθμητικό overflow ή underflow. Χρησιμοποιείται και ένα “ALPHA” επίπεδο για να εξισορροπεί τις νέες λέξεις που εμφανίζονται. Η αξιολόγηση πραγματοποιείται μετρώντας την ορθότητα (accuracy), την ακρίβεια (precision), την ανάκληση (recall) για τα διαφορετικά επίπεδα του ALPHA (εμεις αξιοποιούμαι εδώ την παράμετρο με τιμή 0.01) και σχεδιάζοντας την καμπύλη ακριβείας
(precision) και ανάκλησης (recall). Επιπλέον, εκτυπώνονται οι 20 πιο σημαντικές λέξεις της κάθε τάξης και υλοποιείται μια συνάρτηση που κάνει πρόβλεψη για κάθε κριτική.

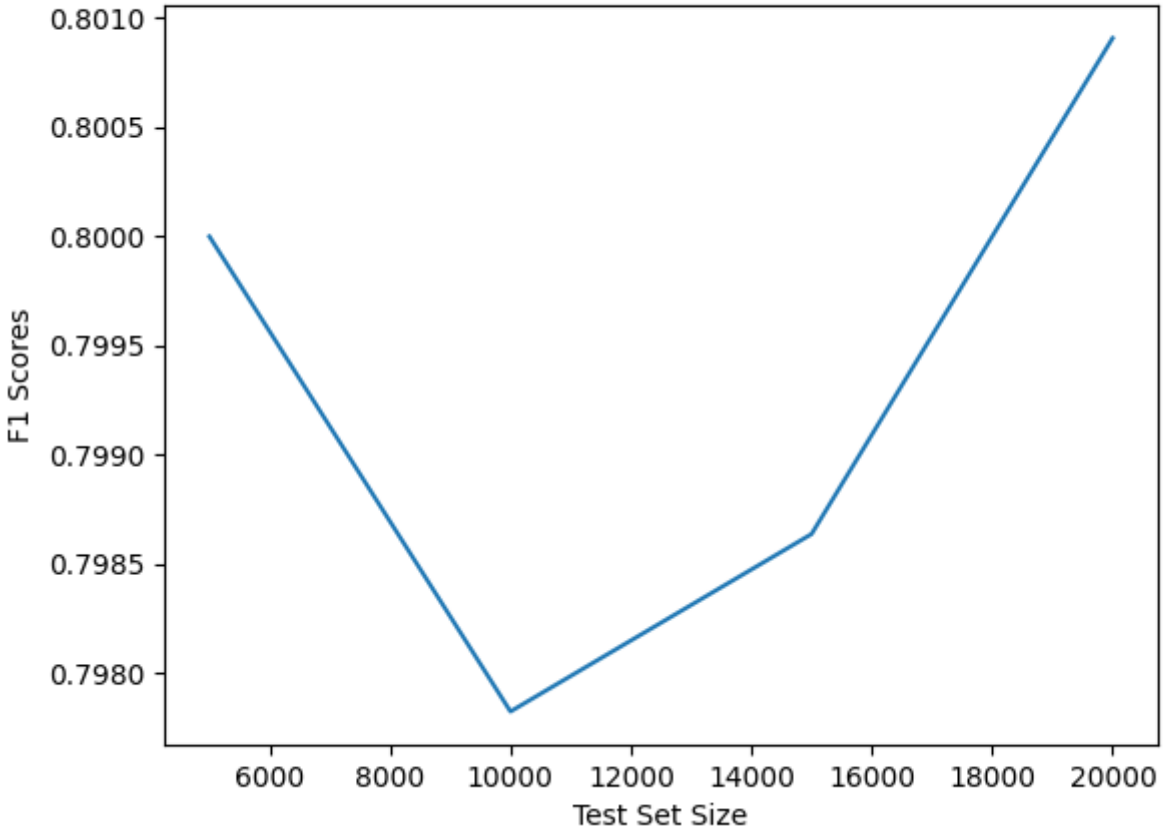
Πιο συγκεκριμένα, χρησιμοποιώντας τα IMDB.py και Vocabulary.py αρχεία παίρνουμε 2 πίνακες σε csr format. Ο πίνακας X είναι ένας μαθηματικός πίνακας που περιγράφει την συχνότητα του κάθε όρου των δεδομένων μας και το Y είναι ένας πίνακας με την ετικέτα (+1 ή -1) της κατηγορίας του κάθε αρχείου . Με τα παραπάνω δημιουργείται ένα λεξικό με το μοναδικά αναγνωριστικά (id) , τους πίνακες συχνοτήτων και τις κατηγορίες αρχείων. Η Train χρησιμοποιείται για τον υπολογισμό απαραίτητων παραμέτρων, όπως οι συνολικές κριτικές, συνολικές θετικές λέξεις, η συχνότητα κάθε λέξης της κατηγορίας, κ.λπ., οι οποίες θα χρησιμοποιηθούν αργότερα για τον υπολογισμό των πιθανοτήτων. Η συνάρτηση Assess χρησιμοποιείται για τη δοκιμή του αλγορίθμου. Καλείται το PredictLabel με τα test data για την αξιολόγηση των προβλέψεων . Η πιθανότητα P(C|W) (P(w δεδομένου c) υπολογίζεται χρησιμοποιώντας λογαριθμικό άθροισμα (έχουμε υλοποιήσει κατάλληλη συνάρτηση) των P(W|C) και P(C). Αυτά με τη σειρά τους έχουν υπολογισθεί από τη συνάρτηση Train. Με βάση το συνολικό θετικό score, το συνολικό αρνητικό score, και ένα δοσμένο όριο (threshold), εκχωρείται μια ετικέτα συγκρίνοντας την πιθανότητα με ένα όριο που έχει καθορισθεί από τον χρήστη. Έπειτα υπολογίζονται τα accuracy, precision, recall ανα αριθμο δεδομένων test ώστε να βλέπουμε την πρόοδο του αλγορίθμου για τις διάφορες μετρικές. Με την συνάρτηση predict_probability αποκτούμε τις ετικέτες που

έχουν
προβλεφθεί, τις πραγματικές ετικέτες και 20 από τα αρχεία από των δεδομένων δοκιμής. Στη συνέχεια λαμβάνουμε τις 20 πιο συχνές θετικές ή αρνητικές λέξεις. Αυτό το κάνουμε υπολογίζοντας το Log odds (odds = ratio of something happening to something not happening) κάθε λέξης σε κάθε τάξη, και δημιουργώντας τη διαφορά αυτών των log odds για τον συνδυασμό κάθε λέξης και κατηγορίας. Υπολογίζουμε δηλαδή τη διαφορά της πιθανότητας P(w|c) των δύο κατηγοριών. Οι εξαιρετικά θετικές τιμές είναι οι λέξεις για τη θετική ετικέτα και οι εξαιρετικά αρνητικές για την αρνητική ετικέτα.




Precission
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size


Recall
0.0
0.2
0.4
0.6
0.8
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size




```
For the positive class:
Calculate the accuracy:  0.7918
Calculate the precision:  0.7607033092702452
Calculate the recall:  0.85144
Calculate the f1 score:  0.8035181759842966


For the negative class:
Calculate the accuracy:  0.7918
Calculate the precision:  0.8313198292306295
Calculate the recall:  0.73216
Calculate the f1 score:  0.7785954315368582
```

```
1.0 Once again Mr. Costner has dragged out a movie for far longer than necessary. Aside from the terrific sea rescue sequences, of which there are very few I just did not care about any of the characters. Most of us have ghosts in the closet, and Costner's charac
ter are realized early on, and then forgotten until much later, by which time I did not care. The character we should really care about is a very cocky, overconfident Ashton Kutcher. The problem is he comes off as kid who thinks he's better than anyone else aroun
d him and shows no signs of a cluttered closet. His only obstacle appears to be winning over Costner. Finally when we are well past the half way point of this stinker, Costner tells us all about Kutcher's ghosts. We are told why Kutcher is driven to be the best w
ith no prior inkling or foreshadowing. No magic here, it was all I could do to keep from turning it off an hour in. -1.0
1.0 This is an example of why the majority of action films are the same. Generic and boring, there's really nothing worth watching here. A complete waste of the then barely-tapped talents of Ice-T and Ice Cube, who've each proven many times over that they are cap
able of acting, and acting well. Don't bother with this one, go see New Jack City, Ricochet or watch New York Undercover for Ice-T, or Boyz n the Hood, Higher Learning or Friday for Ice Cube and see the real deal. Ice-T's horribly cliched dialogue alone makes thi
s film grate at the teeth, and I'm still wondering what the heck Bill Paxton was doing in this film? And why the heck does he always play the exact same character? From Aliens onward, every film I've seen with Bill Paxton has him playing the exact same irritating
 character, and at least in Aliens his character died, which made it somewhat gratifying...<br /><br />Overall, this is second-rate action trash. There are countless better films to see, and if you really want to see this one, watch Judgement Night, which is prac
tically a carbon copy but has better acting and a better script. The only thing that made this at all worth watching was a decent hand on the camera - the cinematography was almost refreshing, which comes close to making up for the horrible film itself - but not
quite. 4/10. 1.0
1.0 First of all I hate those moronic rappers, who could'nt act if they had a gun pressed against their foreheads. All they do is curse and shoot each other and acting like cliché'e version of gangsters.<br /><br />The movie doesn't take more than five minutes to
 explain what is going on before we're already at the warehouse There is not a single sympathetic character in this movie, except for the homeless guy, who is also the only one with half a brain.<br /><br />Bill Paxton and William Sadler are both hill billies and
 Sadlers character is just as much a villain as the gangsters. I did'nt like him right from the start.<br /><br />The movie is filled with pointless violence and Walter Hills specialty: people falling through windows with glass flying everywhere. There is pretty
much no plot and it is a big problem when you root for no-one. Everybody dies, except from Paxton and the homeless guy and everybody get what they deserve.<br /><br />The only two black people that can act is the homeless guy and the junkie but they're actors by
profession, not annoying ugly brain dead rappers.<br /><br />Stay away from this crap and watch 48 hours 1 and 2 instead. At lest they have characters you care about, a sense of humor and nothing but real actors in the cast. -1.0
1.0 Not even the Beatles could write songs everyone liked, and although Walter Hill is no mop-top he's second to none when it comes to thought provoking action movies. The nineties came and social platforms were changing in music and film, the emergence of the Ra
pper turned movie star was in full swing, the acting took a back seat to each man's overpowering regional accent and transparent acting. This was one of the many ice-t movies i saw as a kid and loved, only to watch them later and cringe. Bill Paxton and William S
adler are firemen with basic lives until a burning building tenant about to go up in flames hands over a map with gold implications. I hand it to Walter for quickly and neatly setting up the main characters and location. But i fault everyone involved for turning
 out Lame-o performances. Ice-t and cube must have been red hot at this time, and while I've enjoyed both their careers as rappers, in my opinion they fell flat in this movie. It's about ninety minutes of one guy ridiculously turning his back on the other guy to t
he point you find yourself locked in multiple states of disbelief. Now this is a movie, its not a documentary so i wont waste my time recounting all the stupid plot twists in this movie, but there were many, and they led nowhere. I got the feeling watching this t
hat everyone on set was sord of confused and just playing things off the cuff. There are two things i still enjoy about it, one involves a scene with a needle and the other is Sadler's huge 45 pistol. Bottom line this movie is like domino's pizza. Yeah ill eat it
 if I'm hungry and i don't feel like cooking, But I'm well aware it tastes like crap. 3 stars, meh. -1.0
1.0 Brass pictures (movies is not a fitting word for them) really are somewhat brassy. Their alluring visual qualities are reminiscent of expensive high class TV commercials. But unfortunately Brass pictures are feature films with the pretense of wanting to enter
tain viewers for over two hours! In this they fail miserably, their undeniable, but rather soft and flabby than steamy, erotic qualities non withstanding.<br /><br />Senso '45 is a remake of a film by Luchino Visconti with the same title and Alida Valli and Farle
y Granger in the lead. The original tells a story of senseless love and lust in and around Venice during the Italian wars of independence. Brass moved the action from the 19th into the 20th century, 1945 to be exact, so there are Mussolini murals, men in black sh
irts, German uniforms or the tattered garb of the partisans. But it is just window dressing, the historic context is completely negligible.<br /><br />Anna Galiena plays the attractive aristocratic woman who falls for the amoral SS guy who always puts on too much
 lipstick. She is an attractive, versatile, well trained Italian actress and clearly above the material. Her wide range of facial expressions (signalling boredom, loathing, delight, fear, hate ... and ecstasy) are the best reason to watch this picture and worth t
wo stars. She endures this basically trashy stuff with an astonishing amount of dignity. I wish some really good parts come along for her. She really deserves it. 1.0
1.0 A funny thing happened to me while watching "Mosquito": on the one hand, the hero is a deaf-mute and the director is totally unable to make us understand why he does what he does (mutilating mannequins...er, excuse me, corpses) through his images. On the othe
r hand, the English version at least is very badly dubbed. So I found myself wishing there had been both more AND less dialogue at the same time! This film is stupid (funny how this guy has access to every graveyard and mortuary in his town) and lurid (where woul
d we be in a 70s exploitationer without our gratuitous lesbian scene?). Not to mention the "romantic" aspect (oh, how sweet!)...Miss it. (*) 1.0
1.0 This German horror film has to be one of the weirdest I have seen.<br /><br />I was not aware of any connection between child abuse and vampirism, but this is supposed based upon a true character.<br /><br />Our hero is deaf and mute as a result of repeated b
eatings at the hands of his father. he also has a doll fetish, but I cannot figure out where that came from. His co-workers find out and tease him terribly.<br /><br />During the day a mild-manner accountant, and at night he breaks into cemeteries and funeral hom
es and drinks the blood of dead girls. They are all attractive, of course, else we wouldn't care about the fact that he usually tears their clothing down to the waist. He graduates eventually to actually killing, and that is what gets him caught.<br /><br />Like
 I said, a very strange movie that is dark and very slow as Werner Pochath never talks and just spends his time drinking blood. 1.0
-1.0 Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write to much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they
 called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcedly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this mov
ie was going to move along, and that's not a good thing in my book.<br /><br />Fans of "Beat" Takeshi: his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film.<br /><br />2/10 -1.0
j: awful!   8.156688887052278
j: slater   8.864344021145532
j: drivel.   8.03156494748545
j: segal   8.03156494748545
j: tashan   8.03156494748545
j: kareena   7.997674886166488
j: aztec   7.997674886166488
j: blah,   7.962595877305191
j: />1/10   7.962595877305191
j: sarne   7.926241455837699
j: btk   7.926241455837699
j: delia   7.888515367493735
```

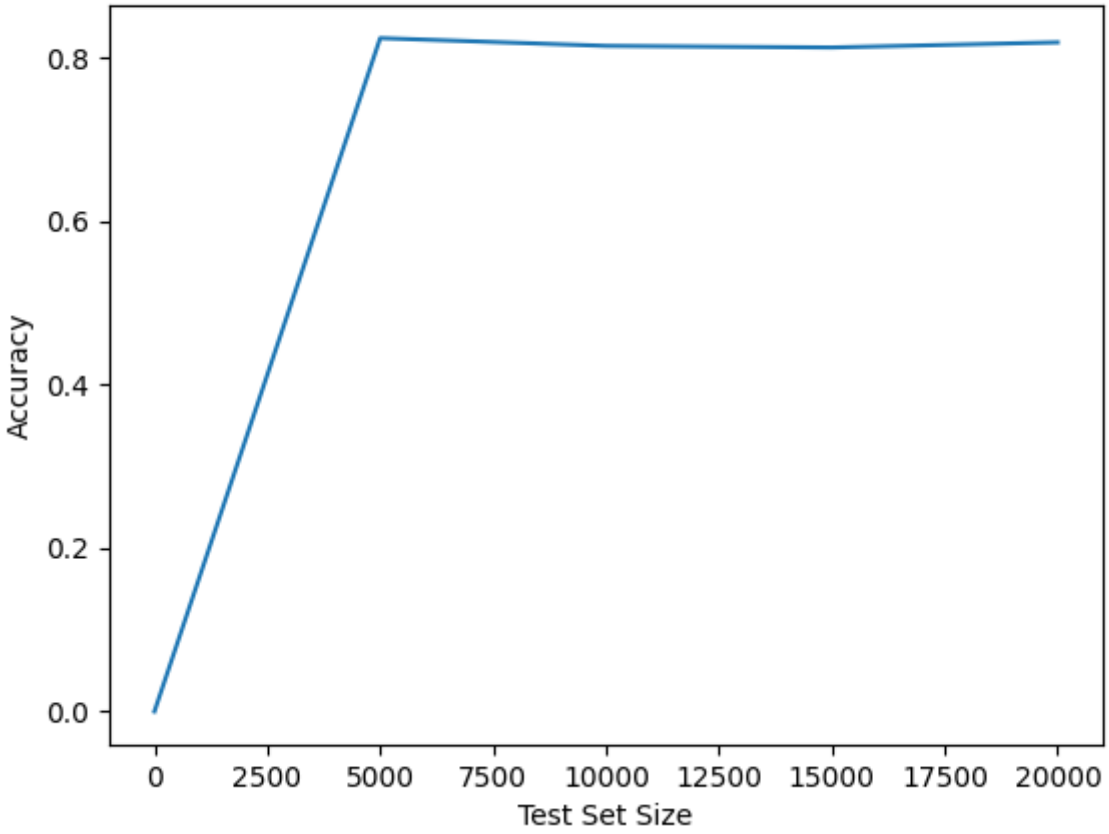
Αλγόριθμος ID3

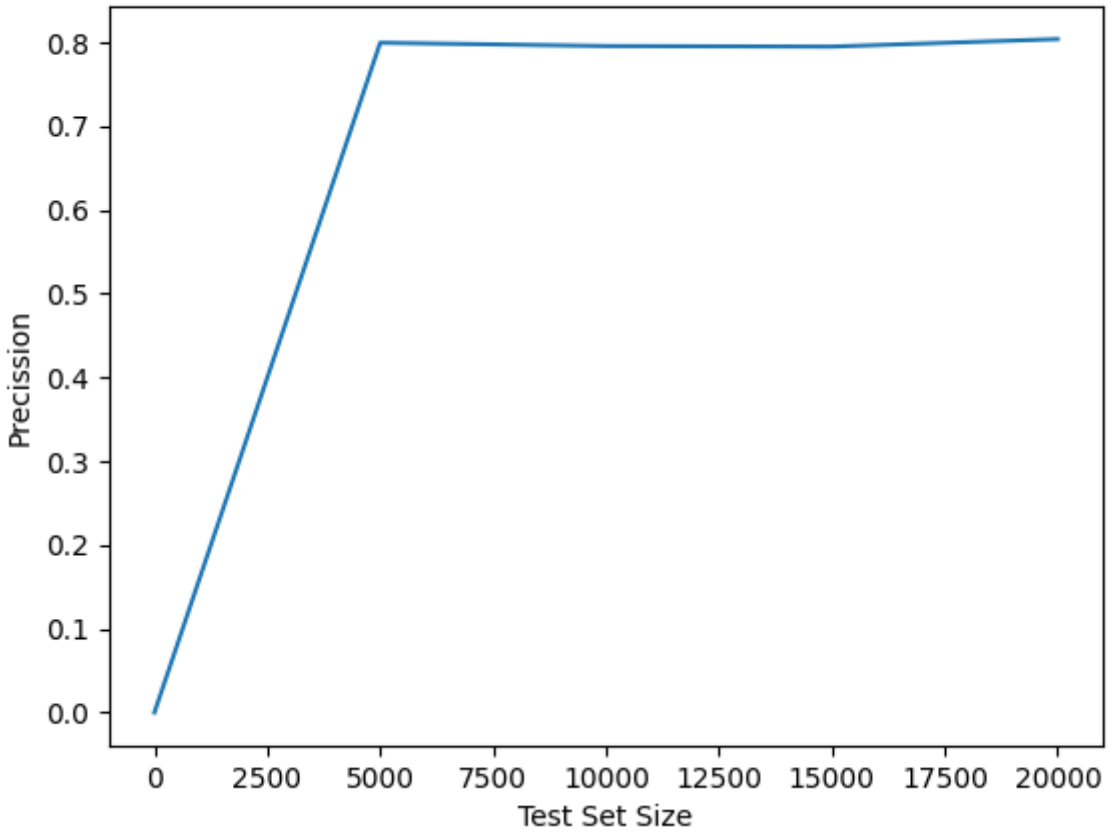
Για τον αλγόριθμο ID3 υλοποιήσαμε 4 κλάσεις. Αυτές είναι οι Node, DecisionTree, ID3, Utility.
Για την εισαγωγή των δεδομένων στον αλγόριθμο ID3 χρησιμοποιήσαμε τις βιβλιοθήκες tensorflow και keras.
ID3: Εδώ έχουμε την κύρια λειτουργία του αλγορίθμου. Στη κλάση αυτή γίνεται η εισαγωγή των δεδομένων από το keras και σε αυτό το αρχείο τρέχει η main. Στην συνάρτηση id3_run παίρνει τα δεδομένα και παίρνει τις παρακάτω περιπτώσεις: να χρησιμοποιήσει το 20%, 40%, 60%, 80% και 100% των δεδομένων εκπαίδευσης. Με βάση αυτά τα δεδομένα φτιάχνει ένα δέντρο αναζήτησης για κάθε ποσοστό δεδομένων. Επίσης καλεί την Utility ώστε να βρει το best_attribute, το information gain και την εντροπία.
DecisionTree: Σε αυτή την κλάση δημιουργείται το δέντρο. Η συνάρτηση add_tree προσθέτει κάποιον κόμβο στο δέντρο ενώ η get_number_of_nodes επιστρέφει τον συνολικό αριθμό των κόμβων. Η κύρια λειτουργία της κλάσης γίνεται στην συνάρτηση classify στην οποία ελέγχει αν τα δεδομένα έχουν διαχωριστεί σωστά σε θετικές και αρνητικές κριτικές.
Node: Η κλάση αυτή ελέγχει αν ο κόμβος είναι φύλλο. Η συνάρτηση get_number επιστρέφει τον αριθμό του κόμβου. Η set_classification, εφόσον ο κόμβος είναι φύλλο, το βάζει σε κατηγορία, δηλαδή σε θετική ή αρνητική κριτική. Η get_classification επιστρέφει την κλάση του κόμβου. Η μέθοδος get_split_value επιστρέφει την τιμή βάσει της οποίας γίνεται και περαιτέρω διαχωρισμός του δέντρου. Η set_node_type ελέγχει αν έχουμε μόνο θετικές ή μόνο αρνητικές, και εφόσον είναι κάνει την μεταβλητή isLeaf True. Η get_majority_classification επιστρέφει αν έχουμε περισσότερες θετικές ή αρνητικές κριτικές.
Utility: Σε αυτή την κλάση, με την μέθοδο get_best_attribute και εφόσον έχουμε υπολογίσει την εντροπία(get_total_entropy) και το information gain(get_information_gain), βρίσκουμε το καλύτερο δεδομένο για να γίνει ο διαχωρισμός

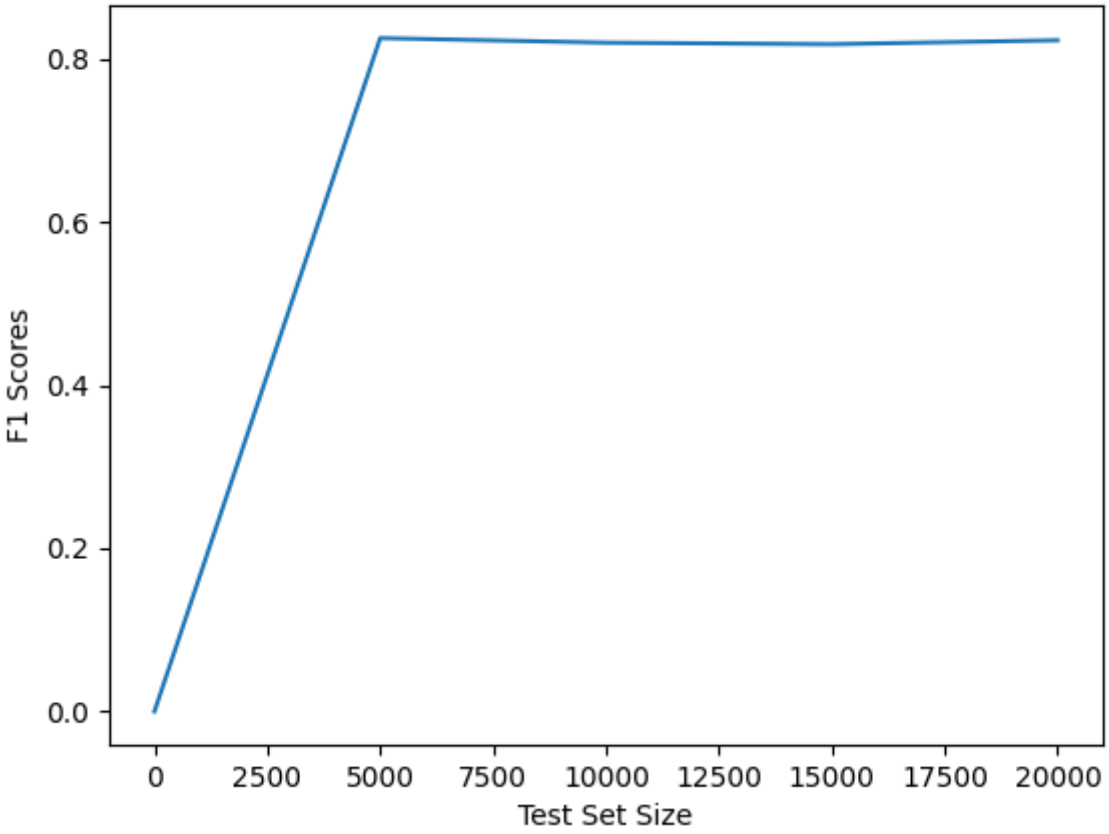
Μέρος Β

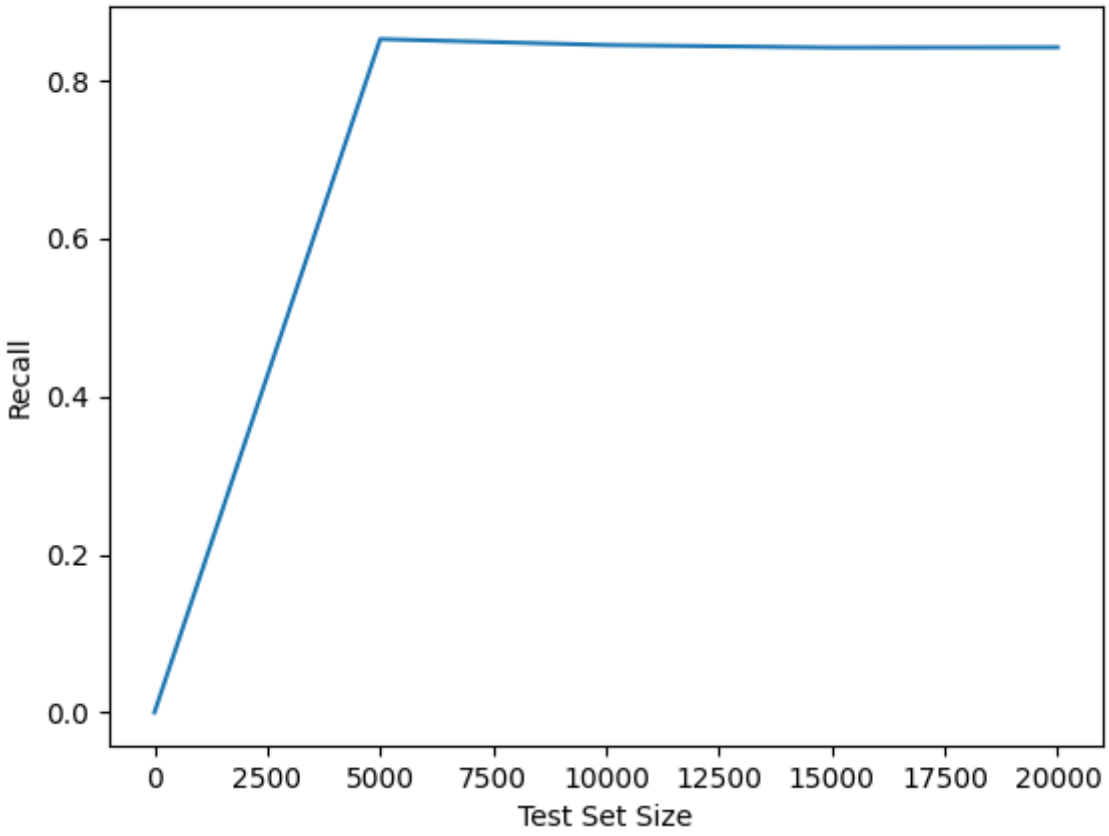
Υλοποίθηκαν

Sklearn_BernoulliNB

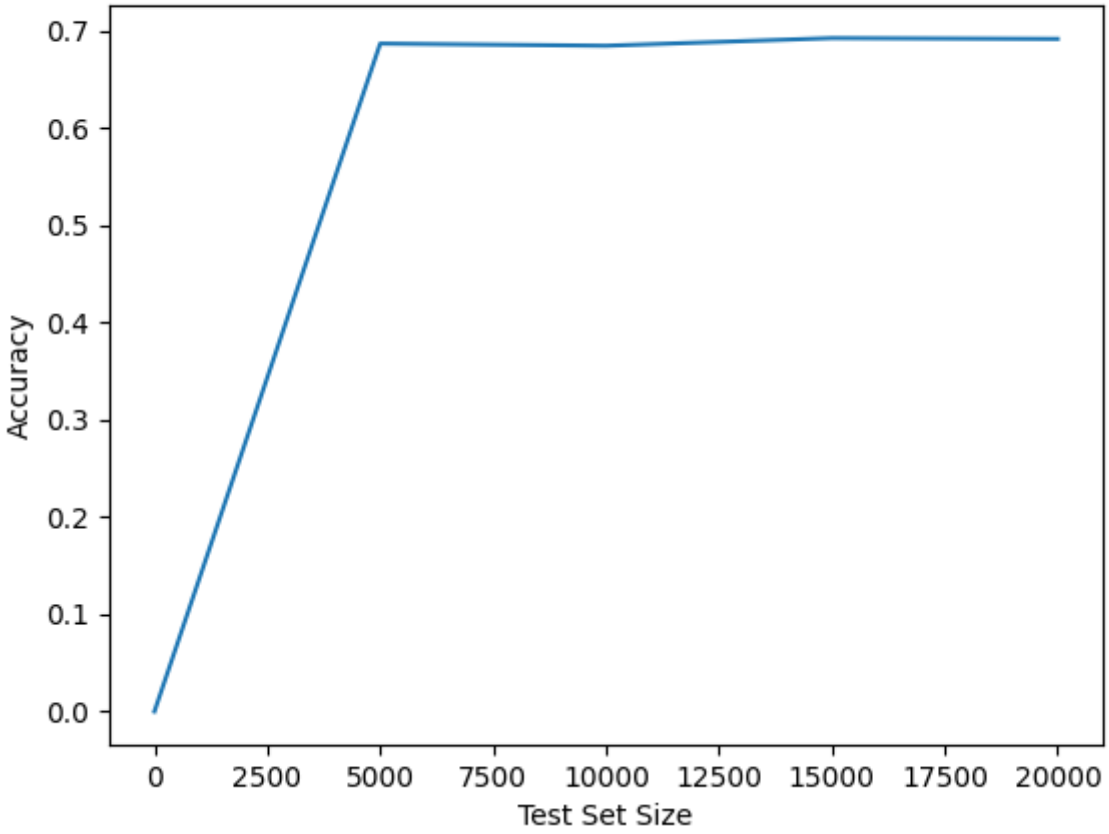

Accuracy
0.0
0.2
0.4
0.6
0.8
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size

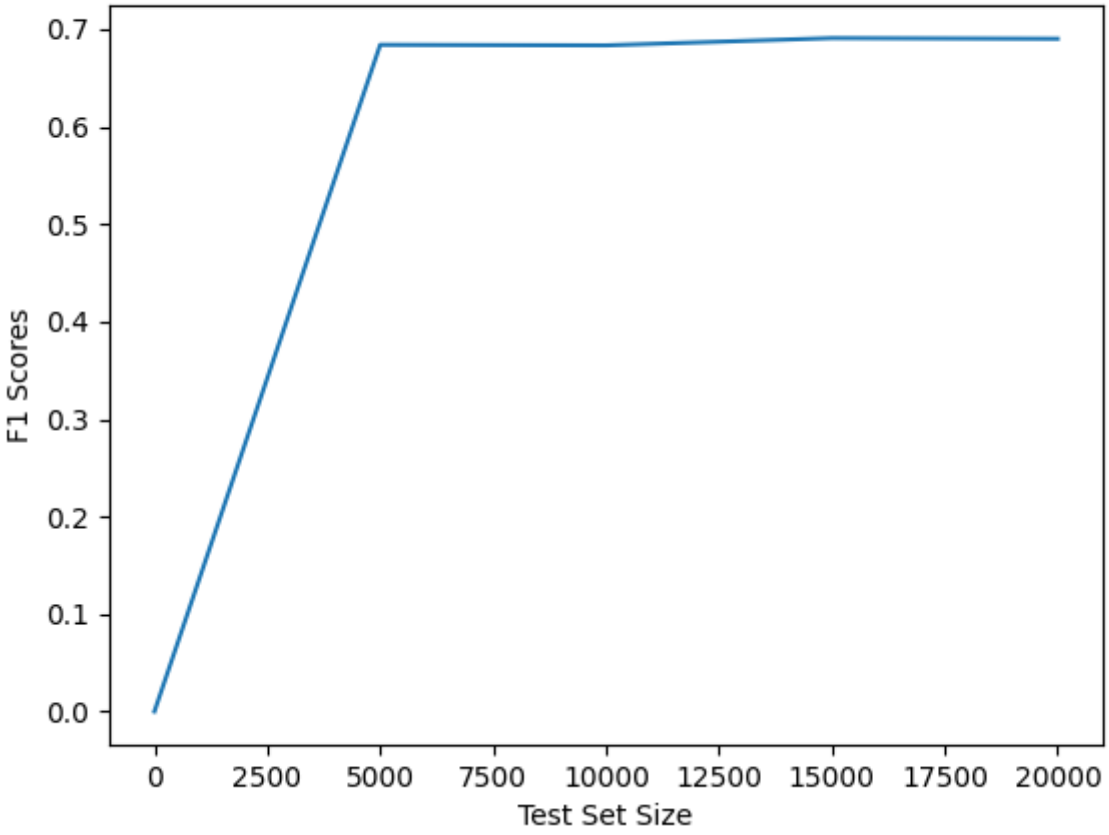

Precission
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size

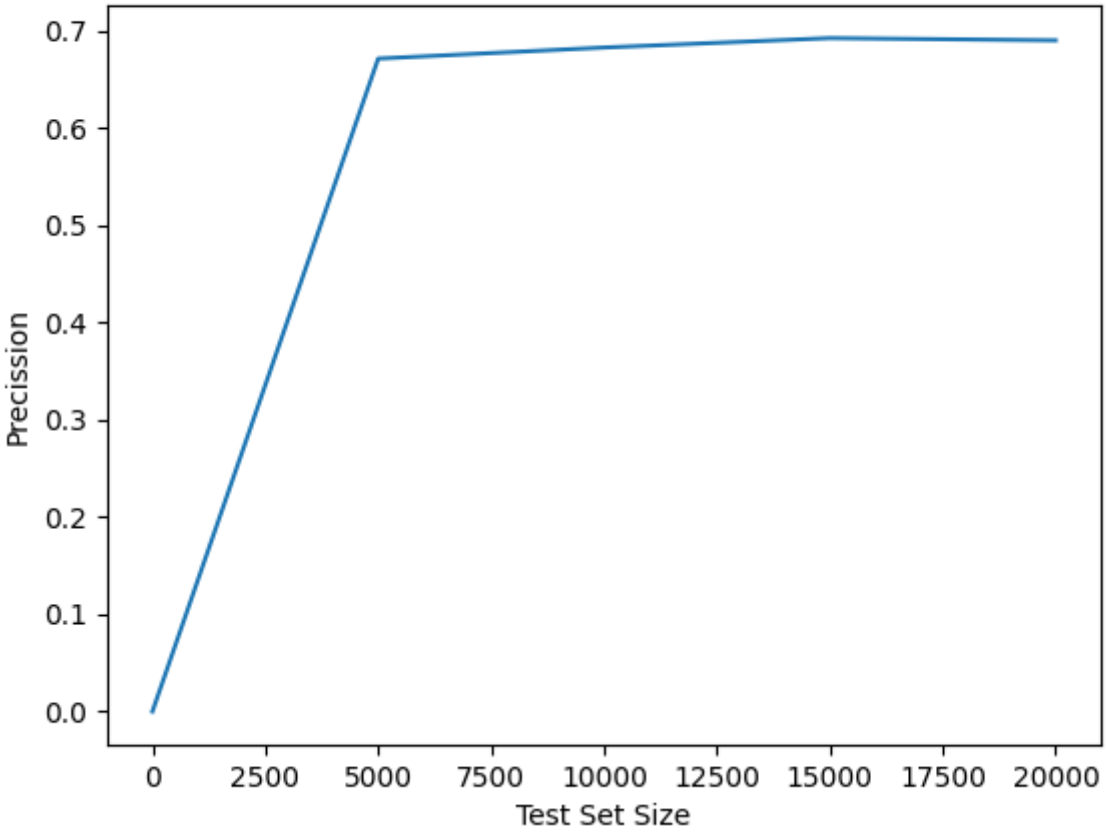


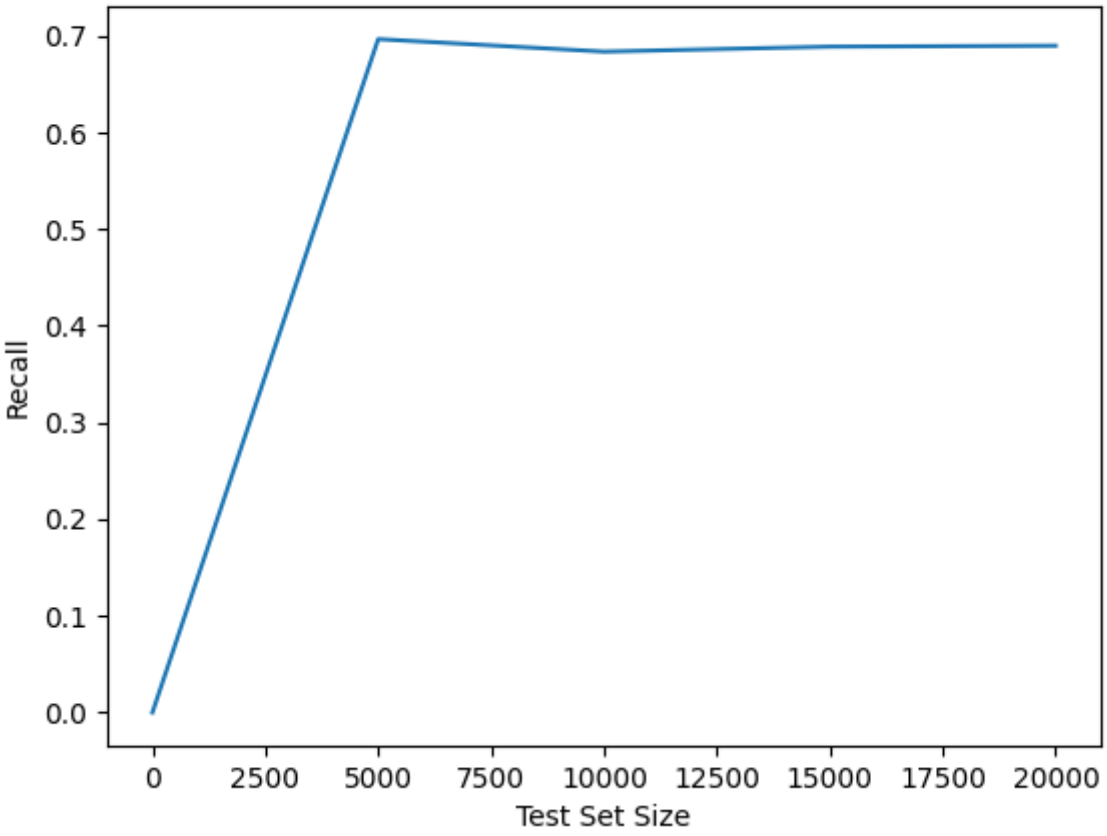


Sklearn_DecissionTreeClassifier


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Accuracy
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
F1 Scores
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size

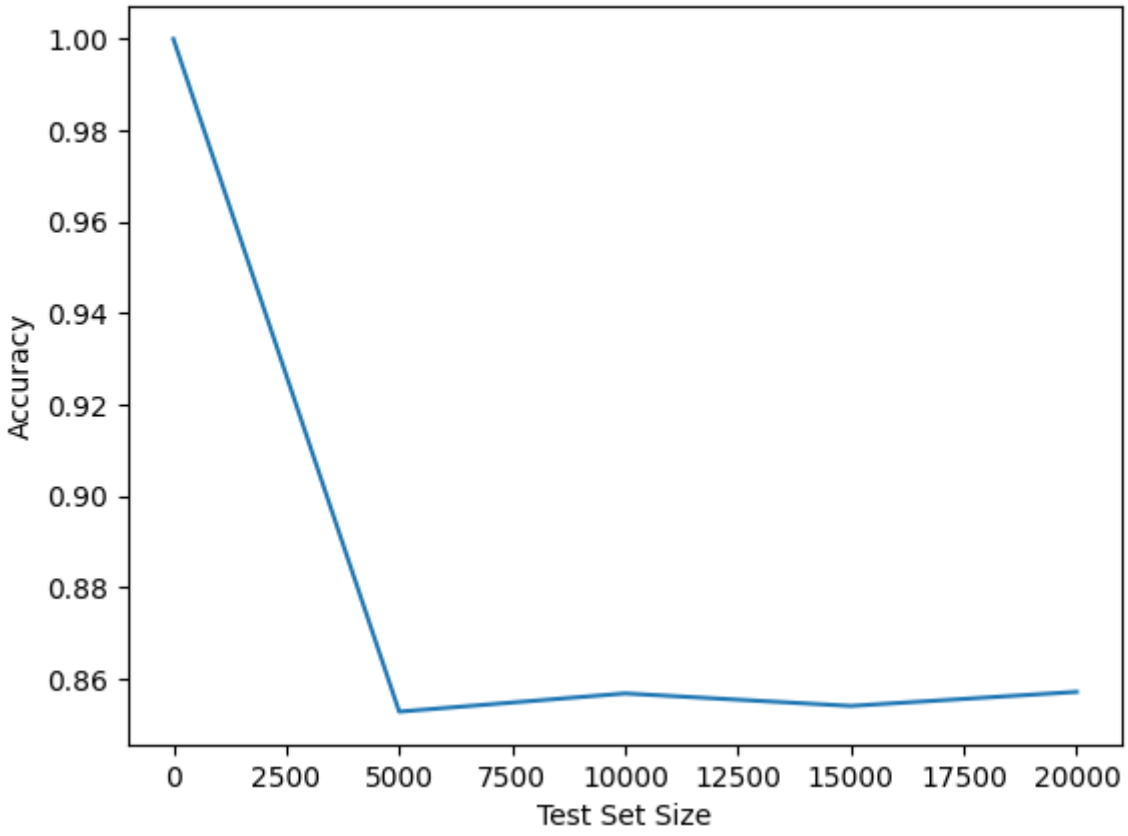

Precission
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size

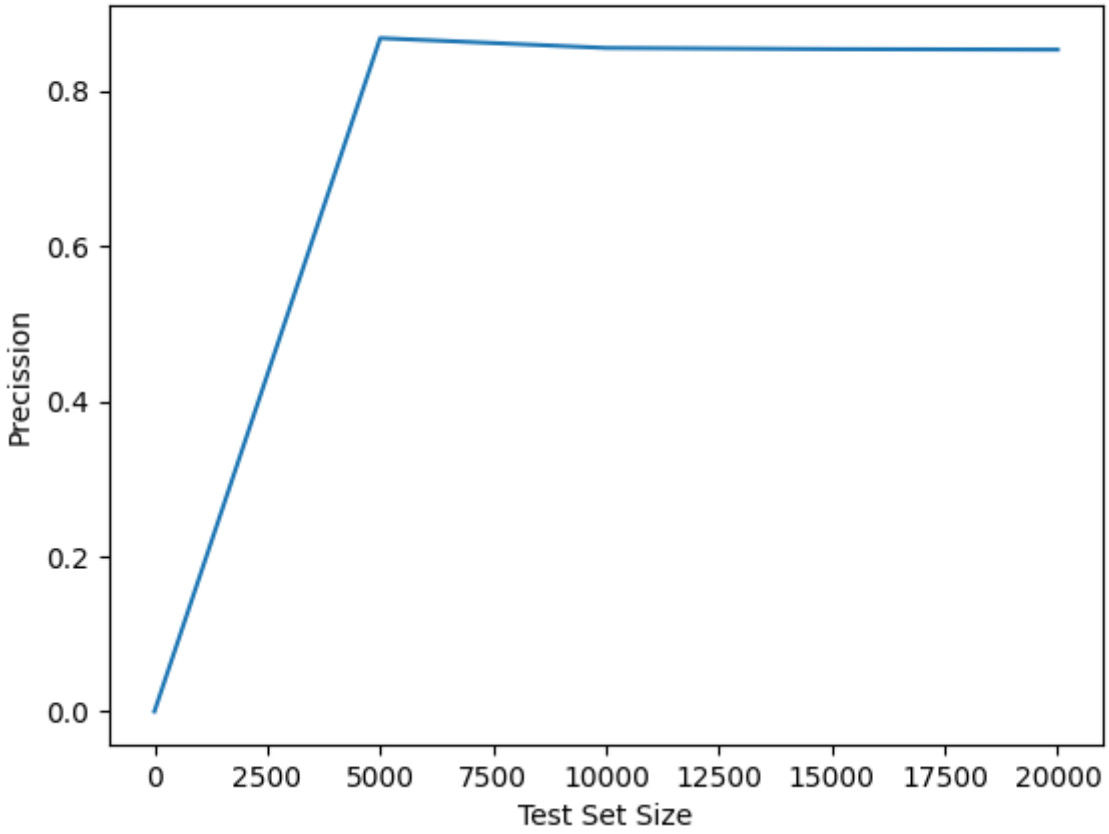

Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size

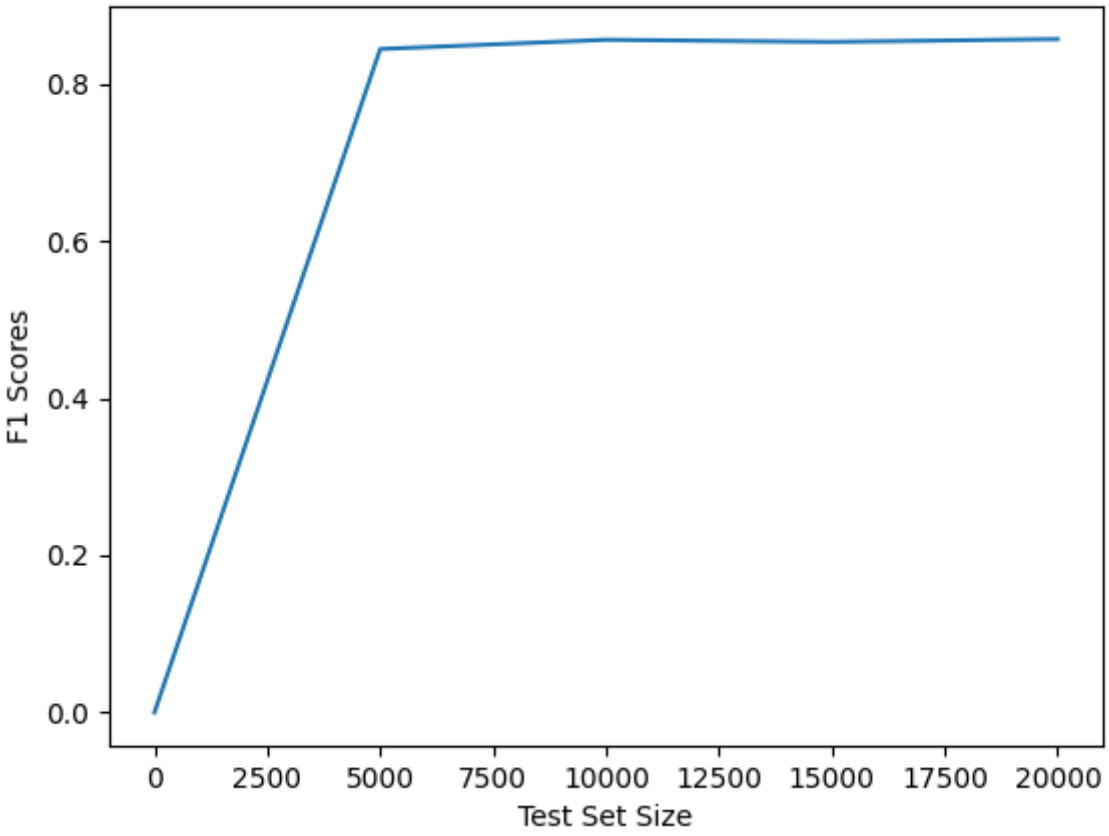
Μέρος Γ

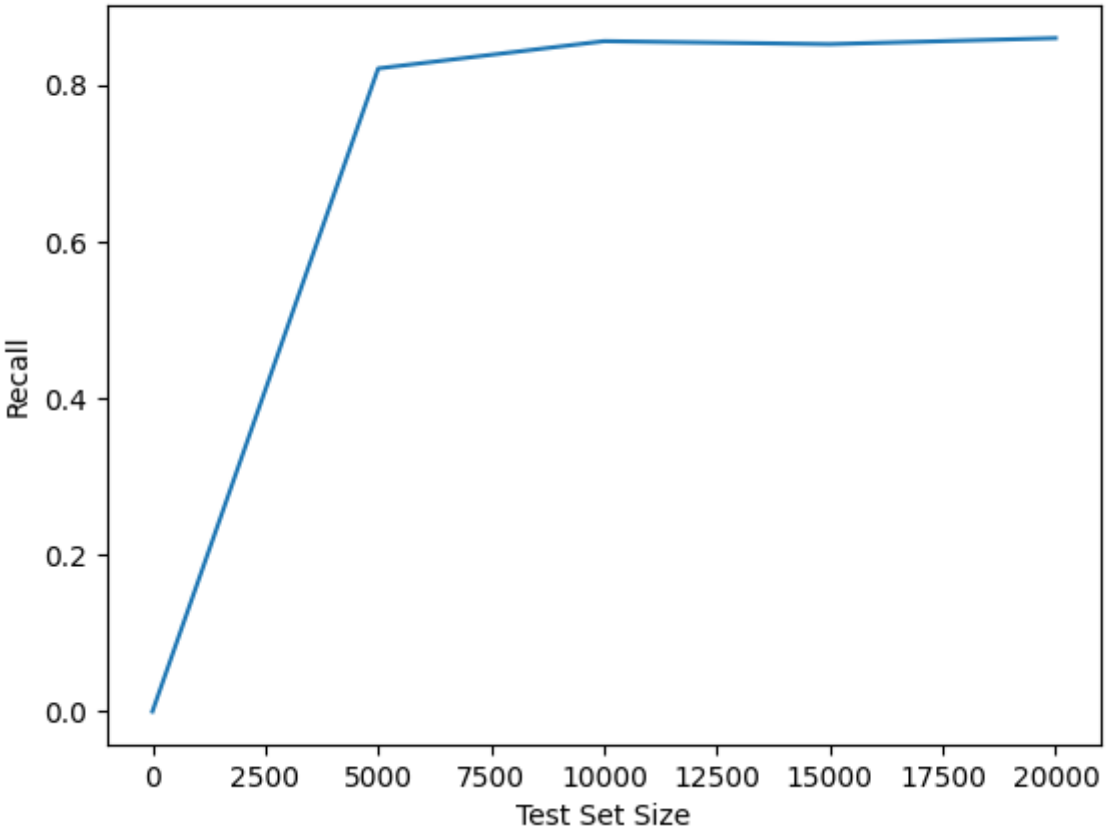
Υλοποιήθηκε στον φάκελο MLP

```
Accuracy: 1.0
1/1 [==============================] - 0s 34ms/step
C:\Python310\lib\site-packages\sklearn\metrics\_classification.py:1344: UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 due to no true samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\Python310\lib\site-packages\sklearn\metrics\_classification.py:1344: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 due to no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\Python310\lib\site-packages\sklearn\metrics\_classification.py:1609: UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 due to no true nor predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, "true nor predicted", "F-score is", len(true_sum))
Epoch 1/4
49/49 [==============================] - 0s 2ms/step - loss: 0.3111 - accuracy: 0.8720
Epoch 2/4
49/49 [==============================] - 0s 2ms/step - loss: 0.3058 - accuracy: 0.8725
Epoch 3/4
49/49 [==============================] - 0s 2ms/step - loss: 0.3027 - accuracy: 0.8739
Epoch 4/4
49/49 [==============================] - 0s 2ms/step - loss: 0.3007 - accuracy: 0.8745
157/157 [==============================] - 0s 581us/step - loss: 0.3372 - accuracy: 0.8528
Accuracy: 0.8528294563293457
157/157 [==============================] - 0s 846us/step
Epoch 1/4
49/49 [==============================] - 0s 2ms/step - loss: 0.3000 - accuracy: 0.8766
Epoch 2/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2958 - accuracy: 0.8759
Epoch 3/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2954 - accuracy: 0.8775
Epoch 4/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2929 - accuracy: 0.8778
313/313 [==============================] - 0s 514us/step - loss: 0.3308 - accuracy: 0.8568
Accuracy: 0.8568143248558044
313/313 [==============================] - 0s 452us/step
Epoch 1/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2895 - accuracy: 0.8790
Epoch 2/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2874 - accuracy: 0.8824
Epoch 3/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2850 - accuracy: 0.8810
Epoch 4/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2815 - accuracy: 0.8826
469/469 [==============================] - 0s 535us/step - loss: 0.3362 - accuracy: 0.8541
Accuracy: 0.8540763854980469
469/469 [==============================] - 0s 458us/step
Epoch 1/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2788 - accuracy: 0.8826
Epoch 2/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2747 - accuracy: 0.8849
Epoch 3/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2707 - accuracy: 0.8872
Epoch 4/4
49/49 [==============================] - 0s 2ms/step - loss: 0.2659 - accuracy: 0.8907
626/626 [==============================] - 0s 518us/step - loss: 0.3299 - accuracy: 0.8572
Accuracy: 0.8571571707725525
626/626 [==============================] - 0s 443us/step
```




Precission
0.0
0.2
0.4
0.6
0.8
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size


F1 Scores
0.0
0.2
0.4
0.6
0.8
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size


Recall
0.0
0.2
0.4
0.6
0.8
0
2500
5000
7500
10000
12500
15000
17500
20000
Test Set Size