# Εργασία Ανάλυσης Δεδομένων στο HIGGS Dataset

With classification algorithms

ΙΑΣΩΝ ΤΖΩΡΤΖΗΣ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΑΘΗΝΑ , ΕΛΛΑΔΑ
iason.tzortzis@gmail.com

**ΠΕΡΙΛΗΨΗ**

Η εργασία αυτή ανατέθηκε σε εμάς από τον καθηγητή Χρήστο Δουλκερίδης με σκοπό την καλύτερη κατανόηση τον αλγόριθμων κατηγοριοποίησης. Έτσι μας ανατέθηκαν συγκεκριμένα dataset ανάλογα με τον αριθμό μητρώου του κάθε φοιτητή .Προσωπικά ,έμενα μου ανατέθηκε το HIGGS dataset που περιείχε ένα μεγάλο πακέτο δεδομένων. Πιο συγκεκριμένα οι πρώτες 21 στήλες αναφέρονται σε κινηματικές ιδιότητες που μετρώνται από τους ανιχνευτές σωματιδίων στον επιταχυντή(2-22).Ενω η πρώτη στήλη αναφέρεται στην ετικέτα της ταξης(1 for signal, 0 for background).Τα τελευταία επτά χαρακτηριστικά (22-28) είναι λειτουργίες των πρώτων 21 χαρακτηριστικών .Η εργασία πραγματοποιήθηκε σε στάδια: Αρχικά χρειάστηκε να πραγματοποιήσουμε την ανάκτηση των δεδομένων έτσι ώστε να μπορέσουμε να τα οπτικοποιήσουμε και να καταλάβουμε τη μορφής είναι τα δεδομένα. Επίσης με την οπτικοποίηση θα καταφέρουμε να κατανοήσουμε αν τα δεδομενα μας θα χρειαστούν προ-επεξεργασία. Δηλαδή άμα υπάρχουν ελλιπείς πληροφορίες που θα πρέπει να συμπληρώσουμε είτε διάφοροι άλλοι παράγοντες που πρέπει να ελέγξουμε ότι τηρούνται. Έπειτα εφόσον έχουμε καθορίσει τι θα γίνει με την προ επεξεργασία θα εφαρμόσουμε τα διάφορα μοντέλα κατηγοριοποίησης που έχουμε διδαχτεί έτσι ώστε να μπορέσουμε να κάνουμε μια πρόβλεψη. Όμως όσο τα διάφορα μοντέλα κατηγοριοποίησης επεξεργάζονται τα δεδομένα εμείς θα χρειαστεί να υπολογίσουμε την ακρίβεια πρόβλεψης του κάθε μοντέλου έτσι ώστε να έχουμε την δυνατότητα να επιλέξουμε το πιο κατάλληλο. Με αποτέλεσμα να πραγματοποιούμε την όσο πιο δυνατόν ακριβής πρόβλεψη.

## 1 ΕΙΣΑΓΩΓΗ

Σκοπός της εργασίας αυτής είναι να εφαρμόσουμε αλγορίθμους κατηγοριοποίησης πάνω σε ένα σύνολο δεδομένων που μας έχει ανατεθεί και να κάνουμε μια πρόβλεψη για τις τιμές της στήλης ,δηλαδή του γνωρίσματος που μας ανατέθηκε μέσα από την εφαρμογή των αλγόριθμων στα σύνολα δεδομένων .Επίσης η εργασία προϋποθέτει την σωστή ανάγνωση, κατανόηση των δεδομένων και την προ επεξεργασία έτσι ώστε να μπορέσουν να επιλεχθούν οι κατάλληλοι αλγόριθμοι κατηγοριοποίησης αλλά και οι κατάλληλοι αλγόριθμοι προσαρμογής, με στόχο την βέλτιστη δυνατή πρόβλεψη των δεδομένων.

### ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Περιγραφή του συνόλου δεδομένων:

Το σύνολο δεδομένων στο όποιο αναφέρομαι παραπάνω είναι το σύνολο δεδομένων 9 με όνομα HIGGS Data Set .Πιο συγκεκριμένα το σύνολο αυτό έχει 28 στήλες με 11000000 καταχωρήσεις/ περιστατικά, οι στήλες περιέχουν 27 διαφορετικά γνωρίσματα και 1 γνώρισμα στόχο. Δηλαδή συνολικά το dataset εχει 28 γνωρίσματα. Όλα τα γνωρίσματα εκτός του γνώρισμα στόχου είναι πραγματικοί αριθμοί ενώ το γνώρισμα στόχου έχει μόνο ακέραιους αριθμούς σαν στήλη και μπορεί να έχει τιμές (0 ή 1).Ενω τα άλλα γνωρίσματα έχουν τιμές από -4 έως 4 συνήθως .Οι ονομασίες των στηλών είναι με τη σειρά από 1 εως 28: class label,lepton pT, lepton eta, lepton phi, missing energy magnitude, missing energy phi, jet 1 pt, jet 1 eta, jet 1 phi, jet

1 b-tag, jet 2 pt, jet 2 eta, jet 2 phi, jet 2 b-tag, jet 3 pt, jet 3 eta, jet 3 phi, jet 3 b-tag, jet 4 pt, jet 4 eta, jet 4 phi, jet 4 b-tag, m_jj, m_jjj, m_lv, m_jlv, m_bb, m_wbb, m_wwbb .

## ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ

Προ-επεξεργασία:

Δεν χρειάστηκε σχεδόν καθόλου προ-επεξεργασία στα δεδομένα λόγο ότι δεν είχε καθόλου κενά και γενικότερα δεν παρουσιάζουν καμία ανάγκη προ-επεξεργασίας .Η μόνη προ-επεξεργασία που πραγματοποίησα ήταν η αφαίρεση του γνωρίσματος στόχου έτσι ώστε να έχω την δυνατότητα να κάνω πρόβλεψη των δεδομένων δίχως την ύπαρξη σφάλματος και την δημιουργία train dataset επειδή ο όγκος των δεδομένων ήταν υπερβολικά μεγάλος .

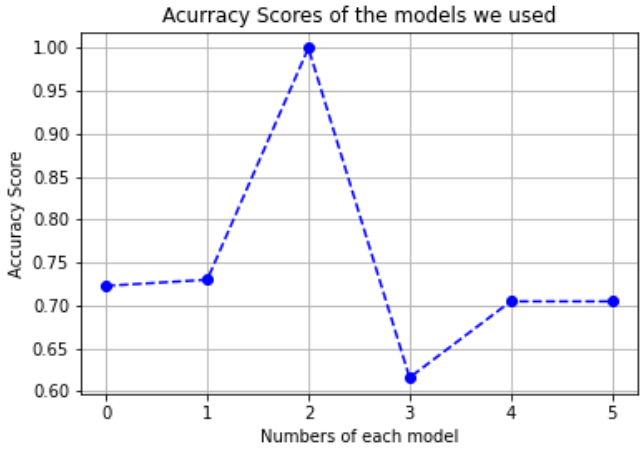
## ΜΕΘΟΔΟΛΟΓΙΑ

Μεθοδολογία :

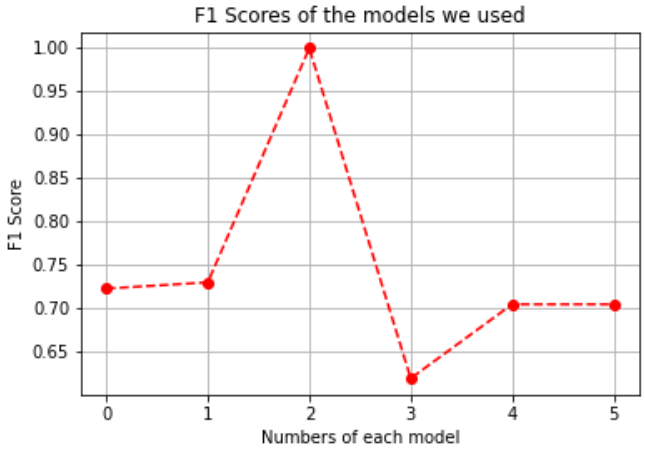
Στο πρόγραμμα αρχικά χρησιμοποιείται μια μέθοδος με την όποια ανακτάται το σύνολο δεδομένων στο data ,όπου στην ίδια μέθοδο αποθηκεύονται και τα ονόματα των γνωρισμάτων (columns) .Έπειτα δημιουργείται το σύνολο δεδομένων που θα χρησιμοποιηθεί για δοκιμή (train_set) και το σύνολο δεδομένων που θα χρησιμοποιηθούν για έλεγχο (test_set) .Καθώς το αρχικό σύνολο δεδομένων είναι υπερβολικά ογκώδες επειδή έχει 11000000 καταχωρήσεις/ περιστατικά, επιδέχθηκε το σύνολο δεδομένων δοκιμής να καταχωρηθεί το 0.001% του αρχικού .Θα ήθελα να χρησιμοποιήσω και αλλά δεδομένα αλλά με την παραμικρή αύξηση το πρόγραμμα δεν θα καταφέρνει ποτέ να τελειώσει ακόμα και μετά από 2 ώρες επεξεργασίας .Εφόσον έχουμε χωρίσει κύριο σύνολο δεδομένων σε σύνολο δεδομένων δοκιμής και έλεγχου μπορούμε να χωρίσουμε το σύνολο δεδομένων δοκιμής περαιτέρω δημιουργώντας έτσι μια νέα λίστα (class label),η ποια αποθηκεύει την στήλη του γνώρισμα στόχου και άλλη μια στήλη με τα predictors.Από αυτό το σημείο και μετά το πρόγραμμα είναι έτοιμο έτσι ώστε να του εφαρμοστούν οι αλγόριθμοι κατηγοριοποίησης. Αρχικά χρησιμοποιείται ο αλγόριθμος SVM. Έπειτα ο αλγόριθμος K-nearest neighbor.Μετέπειτα γίνεται χρήση του αλγορίθμου Decision Tree Classifier.Παρατηρώ ότι αυτός ο αλγόριθμος παρουσιάζει τιμή Accuracy=1.0, πράγμα που σημαίνει ότι ο αλγόριθμος αυτός παρουσιάζει υπερπροσαρμογή στα δεδομένα. Για να το αντιμετωπίσουμε αυτό, εφαρμόζουμε πάλι τον ίδιο αλγόριθμο αλλά αυτή την φορά με διασταυρωτική επικύρωση. Τέλος εφαρμόζεται ο αλγόριθμος Random forest Classifier ο ποιος φαίνεται να εμφανίζει τα καλυτέρα αποτελέσματα από όλους τους προηγούμενους με αποτέλεσμα οι αλγόριθμοι προσαρμογής να εφαρμόζονται πάνω σε αυτόν τον αλγόριθμο και έπειτα από διαφορές δοκιμές το πρόγραμμα βρίσκει τον αλγόριθμο με τις βέλτιστες παραμέτρους και τον εφαρμόζει στο σύνολο δεδομένων για να πραγματοποίηση έλεγχο της απόδοσης του αλγορίθμου.

ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ:
Για την διαδικασία της πειραματικής αξιολόγησης των μοντέλων κατηγοριοποίησης που χρησιμοποιήθηκαν , θα χρησιμοποιήσουμε τις μετρικές Accuracy και F1 Score. Παρακάτω απεικονίζονται τα διαγράμματα που συγκρίνουν τα αποτελέσματα Accuracy και F1 Score αντίστοιχα, που προέκυψαν από κάθε μοντέλο που χρησιμοποιήθηκε . Πιο συγκεκριμένα στον άξονα X υπάρχουν οι τιμές που κυμαίνονται από 0-5 που απεικονίζουν τα 6 αποτελέσματα που προέκυψαν. Ο κάθε αριθμός αντιστοιχεί στο εξής μοντέλο:
0: Αλγόριθμος Support Vector Machines
1: Αλγόριθμος K-nearest neighbor
2: Αλγόριθμος Decision Tree Classifier - Χωρίς Cross Validate
3: Αλγόριθμος Decision Tree Classifier – Με Cross Validate
4: Αλγόριθμος Random Forest Classifier
5: Αλγόριθμος Random Forest Classifier – Μοντέλο με τις βέλτιστες παραμέτρους

## Διάγραμμα Accuracy:




## Διάγραμμα F1 Score:




## Συμπεράσματα:

**Από τα αποτελέσματα τις εργασίας μπορώ να συμπεράνω ότι ο όγκος των δεδομένων που αναλύθηκε δεν κατάφερε να δώσει μια σαφή εικόνα για το γνώρισμα στόχο ως προς την ακρίβεια της πρόβλεψης των τιμών του γνώρισμα στόχου. Δηλαδή ακόμα και ο πιο παραγωγικός αλγόριθμος με την βοήθεια αλγόριθμων προσαρμογής και με τις βέλτιστες παραμέτρους. Για αυτό το γεγονός φταίει ότι ανέλυσα ένα πολύ μικρό μέρος των δεδομένων λόγο του τεράστιου όγκου τους και λόγο του limitation του Google Collaboratory.**

## Βιβλιογραφικές πηγές:

1) Διαφάνειες μαθήματος «Ανάλυση Δεδομένων», Πανεπιστήμιο Πειραιώς, Τμήμα Ψηφιακών Συστημάτων

2) Διαφάνειες Εργαστηρίου μαθήματος «Ανάλυση Δεδομένων», Πανεπιστήμιο Πειραιώς, Τμήμα Ψηφιακών Συστημάτων

3) P. Tan, M. Steinbach, V. Kumar. "Εισαγωγή στην Εξόρυξη Δεδομένων". ΕΚΔΟΣΕΙΣ ΤΖΙΟΛΑ