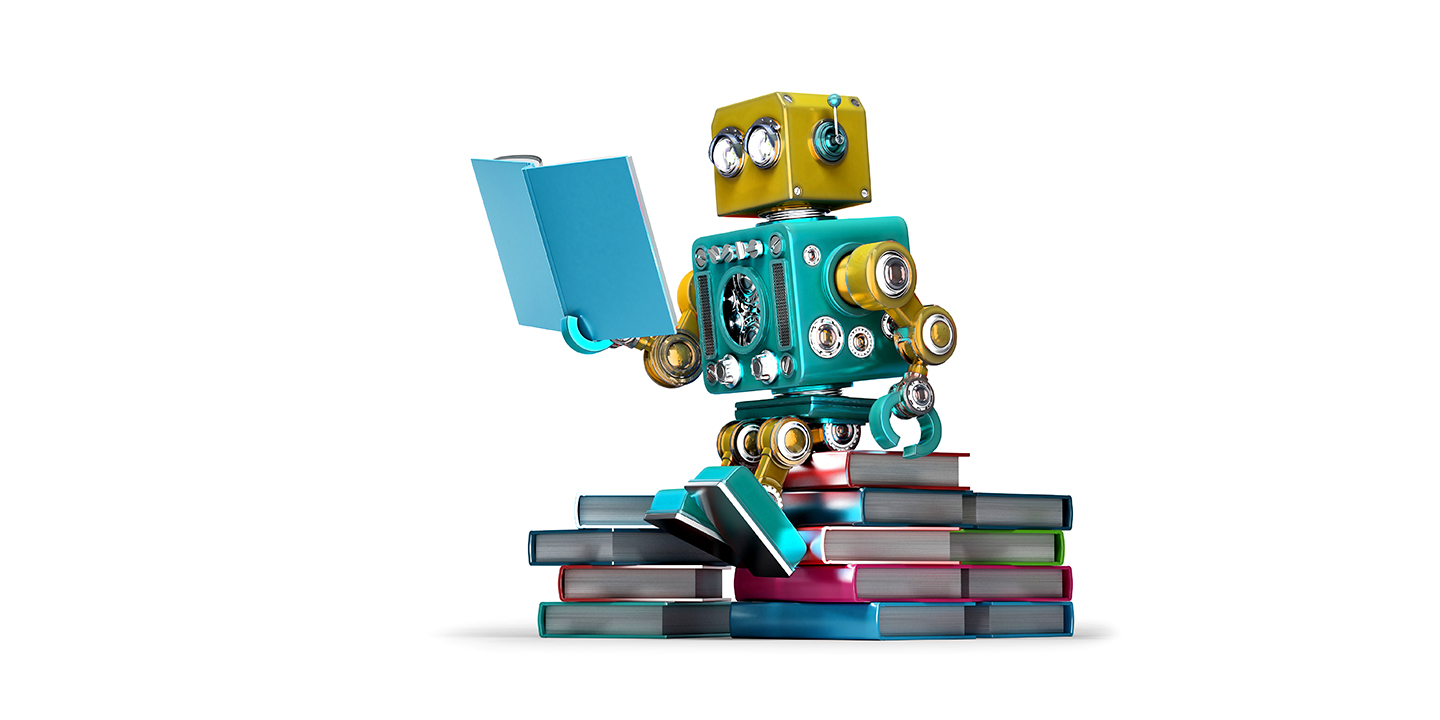
**Μηχανική Μάθηση Εργασία 1**



**Μπόζας Αριστείδης**

**ΑΜ:740**

**Δερμεντζόγλου Ιωάννης**

**ΑΜ:743**

**Περιεχόμενα**

[1. Μέρος Α 3](#_Toc513202900)

[2. Μέρος Β 5](#_Toc513202901)

[3. Μέρος Γ 6](#_Toc513202902)

# Μέρος Α

Στο πρώτο μέρος της εργασίας εξετάστηκε το αντικείμενο των τεχνικών πολλαπλών μοντέλων πρόβλεψης σε συνδυασμό με το αντικείμενο της συγκριτικής αξιολόγησης μεταξύ αλγορίθμων.

Τα 10 datasets που χρησιμοποιήθηκαν από το [UCI](http://archive.ics.uci.edu/ml/index.php) repository είναι τα εξής:

* [Iris](https://archive.ics.uci.edu/ml/datasets/iris)
* [Wine](https://archive.ics.uci.edu/ml/datasets/wine)
* [Breast Cancer Wisconsin (Diagnostic)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
* [Balance Scale](http://archive.ics.uci.edu/ml/datasets/balance+scale)
* [Hayes-Roth](https://archive.ics.uci.edu/ml/datasets/Hayes-Roth)
* [Haberman’s Survival](https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival)
* [Liver Disorder](https://archive.ics.uci.edu/ml/datasets/liver+disorders)s
* [Banknote Authentication](https://archive.ics.uci.edu/ml/datasets/banknote+authentication)
* [Ionosphere](https://archive.ics.uci.edu/ml/datasets/ionosphere)
* [Contraceptive Method Choice](https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice)

Οι τεχνικές που χρησιμοποιήθηκαν για τα πολλαπλά μοντέλα πρόβλεψης είναι οι εξής:

* **Manipulating the training examples:** Oι τεχνικές του bagging και boosting.
* **Manipulating the target variable:** Oι τεχνικές του OnevsOne και OnevsRest.
* **Injecting randomness :** Tο ensemble μοντέλο RandomForest.
* **Manipulating Features:** Τυχαία επιλογή των features και των παραδειγμάτων εκπαίδευσης με την τεχνική RandomPatches.

Όσον αφορά στα αποτελέσματα της συγκριτικής αξιολόγησης μεταξύ των αλγορίθμων που εκτελέστηκαν, αυτά παρουσιάζονται παρακάτω:

Ο πίνακας με τις τιμές της μετρικής accuracy έχει ως εξής:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **DT** | **Bagged DT** | **AdaBoost DT** | **GradientBoost DT** | **OneVsOne DT** | **OneVsRest DT** | **RF** | **Bagged DT RP** |
| **Iris** | 0.96 | 0.967 | 0.953 | 0.96 | 0.953 | 0.96 | 0.96 | 0.96 |
| **Wine** | 0.895 | 0.967 | 0.9 | 0.917 | 0.9 | 0.878 | 0.978 | 0.973 |
| **Breast Cancer** | 0.918 | 0.958 | 0.914 | 0.961 | 0.918 | 0.918 | 0.965 | 0.958 |
| **Balance Scale** | 0.68 | 0.679 | 0.667 | 0.705 | 0.635 | 0.743 | 0.683 | 0.602 |
| **Hayes-Roth** | 0.811 | 0.826 | 0.772 | 0.772 | 0.811 | 0.779 | 0.825 | 0.811 |
| **Haberman's Survival** | 0.626 | 0.638 | 0.569 | 0.59 | 0.619 | 0.626 | 0.671 | 0.655 |
| **Liver Disorders** | 0.567 | 0.58 | 0.568 | 0.58 | 0.567 | 0.567 | 0.571 | 0.566 |
| **Banknote Authentication** | 0.983 | 0.991 | 0.983 | 0.996 | 0.983 | 0.983 | 0.993 | 0.977 |
| **Ionosphere** | 0.881 | 0.918 | 0.879 | 0.93 | 0.881 | 0.881 | 0.938 | 0.932 |
| **Contraceptive Method Choice** | 0.468 | 0.522 | 0.5 | 0.57 | 0.479 | 0.467 | 0.511 | 0.518 |

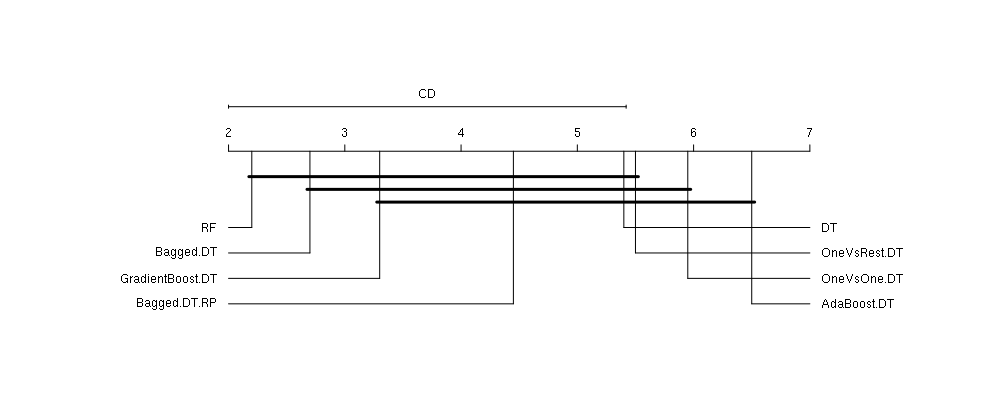
Από αυτόν τον πίνακα προκύπτει ο πίνακας με τις μέσες κατατάξεις των μοντέλων:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **DT** | **Bagged DT** | **AdaBoost DT** | **GradientBoost DT** | **OneVsOne DT** | **OneVsRest DT** | **RF** | **Bagged DT RP** |
| **Iris** | 4 | 1 | 7.5 | 4 | 7.5 | 4 | 4 | 4 |
| **Wine** | 7 | 3 | 5.5 | 4 | 5.5 | 8 | 1 | 2 |
| **Breast Cancer** | 6 | 3.5 | 8 | 2 | 6 | 6 | 1 | 3.5 |
| **Balance Scale** | 4 | 5 | 6 | 2 | 7 | 1 | 3 | 8 |
| **Hayes-Roth** | 4 | 1 | 7.5 | 7.5 | 4 | 6 | 2 | 4 |
| **Haberman's Survival** | 4.5 | 3 | 8 | 7 | 6 | 4.5 | 1 | 2 |
| **Liver Disorders** | 6 | 1.5 | 4 | 1.5 | 6 | 6 | 3 | 8 |
| **Banknote Authentication** | 5.5 | 3 | 5.5 | 1 | 5.5 | 5.5 | 2 | 8 |
| **Ionosphere** | 6 | 4 | 8 | 3 | 6 | 6 | 1 | 2 |
| **Contraceptive Method Choice** | 7 | 2 | 5 | 1 | 6 | 8 | 4 | 3 |
| **Average Ranks** | **5.4** | **2.7** | **6.5** | **3.3** | **5.95** | **5.5** | **2.2** | **4.45** |

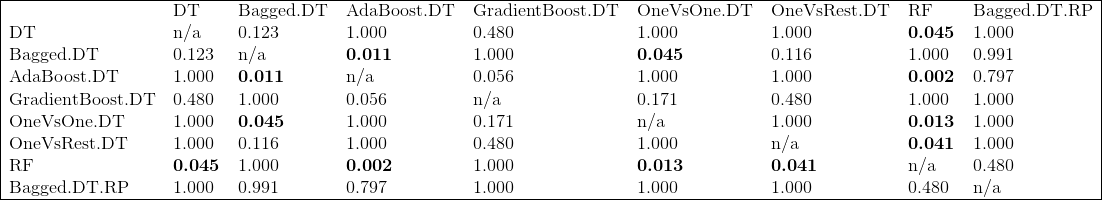
Βασισμένοι στα προηγούμενα μπορούμε να κατατάξουμε τα μοντέλα από το καλύτερο στο χειρότερο ως εξής: RF, Bagged DT, GradientBoost DT, Bagged DT RP, DT, OneVsRest DT, OneVsOne DT, AdaBoost DT.

Για την στατιστική ανάλυση χρησιμοποιούμε το Iman - Davenport τεστ για την εύρεση στατιστικά σημαντικών διαφορών στην απόδοση των μοντέλων, λαμβάνουμε p-value < 0.0001 και συμπεραίνουμε ότι υπάρχουν στατιστικά σημαντικές διαφορές.

Κάνοντας χρήση του Nemenyi τεστ λαμβάνουμε αποτελέσματα που συνοψίζονται στο ακόλουθο διάγραμμα:



από το οποίο συμπεραίνουμε πως το RF έχει στατιστικά σημαντικές διαφορές από τα DT, OneVsRest DT, OneVsOne DT και AdaBoostDT και το Bagged DT από τα OneVsOne DT και AdaBoost DT. Τα συμπεράσματα αυτά επαληθεύονται και από την εφαρμογή ενός Friedman post-hoc τεστ με διόρθωση Bergmann and Hommel των οποία τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα:



# Μέρος Β

Το δεύτερο μέρος της εργασία εξετάστηκε το πρόβλημα του διαφορετικού κόστους στο σύνολο δεδομένου [heart](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)) .H βιβλιοθήκη που χρησιμοποιήθηκε ήταν η **Costcla,sklearn** της python.

To cost matrix αυτού του συνόλου δεδομένου είναι το εξής:

|  |  |  |
| --- | --- | --- |
|  | **Actual absence** | **Actual presence** |
| **Absence** | 0 | 1 |
| **Presence** | 5 | 0 |

Πίνακας 2: Cost matrix του συνόλου δεδομένου heart

Οι μέθοδοι που χρησιμοποιήθηκαν και αναφέρονται παρακάτω συνδιάστηκαν με τους αλγορίθμους μάθησης Random Forest, Linear SVM,Naive Bayes:

* Δίνοντας έμφαση στα παραδείγματα με το μεγαλύτερο κόστος(**CostSampling** [Oversampling, RejectionSampling) )
* Ελαχιστοποίηση αναμενόμενου κόστους εκτιμήσεων (**ThresholdOptimization,BayesMinimumRiskClassifier)**
* Τροποποιημένη cost sensitive ταξινομητές( **CostSensitiveRandoForestClassifier**)

Επειδή οι πιθανότητες παίζουν μεγάλο ρόλο στο πρόβλημα του διαφορετικού κόστους έγινε ένα calibration στις πιθανότητες των αλγορίθμων μάθησης Random Forest, Linear SVM,Naive Bayes με τη χρήση της βιβλιοθήκης **sklearn.calibration.**

|  |  |
| --- | --- |
| **Αλγόριθμος** | **Costloss** |
| **RF** | 59.0 |
| **RF - O** | 41.0 |
| **RF - R** | 41.0 |
| **RF - U** | 53.0 |
| **LSVM** | 60.0 |
| **LSVM - O** | 42.0 |
| **LSVM - R** | 37.0 |
| **LSVM - U** | 61.0 |
| **GNB** | 54.0 |
| **GNB - O** | 45.0 |
| **GNB – R** | 32.0 |
| **GNB - O** | 50.0 |
| **RF - BMR** | 28.0 |
| **RF - TO** | 165.0 |
| **LSVM - BMR** | 33.0 |
| **LSVM - TO** | 37.0 |
| **GNB - BMR** | 33.0 |
| **GNB - TO** | 33.0 |
| **RFC** | 43 |

# Μέρος Γ

Στο τρίτο μέρος της εργασίας εξετάστηκε το πρόβλημα της ασυμμετρίας κλάσεων στο σύνολο δεδομένων [creditcardfraud](https://www.kaggle.com/mlg-ulb/creditcardfraud) που περιέχει 284807 συναλλαγές με πιστωτικές κάρτες, στο οποίο μόνο το 0.172% αυτών είναι απάτες (θετική κλάση). Η ανάλυση πραγματοποιήθηκε με τη βοήθεια της βιβλιοθήκης **imblearn** της python.

Οι μέθοδοι που χρησιμοποιήθηκαν και αναφέρονται παρακάτω συνδιάστηκαν με τους αλγορίθμους μάθησης του **scikit-learn** RandomForestClassifier, LinearSVC και GaussianNB:

* **Over-sampling:** SMOTE
* **Under-sampling:** NearMiss (version=2)
* **Ensemble of samplers:** EasyEnsemble μέσω του BalancedBaggingClassifier

Στα αποτελέσματα που δίνονται παρακάτω χρησιμοποιούνται οι μετρικές precision(pre), recall(rec), specificity(spe), geometric mean(geo), και index balanced accuracy of the geometric mean(iba) για τις δύο κλάσεις.

|  |
| --- |
| **NearMiss-2 - RandomForestClassifier** |
| **pre rec spe f1 geo iba sup**  0.998 0.008 0.992 0.016 0.089 0.007 71079 0.002 0.992 0.008 0.003 0.089 0.009 123  0.997 0.010 0.990 0.016 0.089 0.007 71202 |

|  |
| --- |
| **NearMiss-2 - LinearSVC** |
| **pre rec spe f1 geo iba sup**  1.000 0.173 0.967 0.295 0.409 0.154 71079 0.002 0.967 0.173 0.004 0.409 0.181 123  0.998 0.175 0.966 0.295 0.409 0.154 71202 |

|  |
| --- |
| **NearMiss-2 - GaussianNB** |
| **pre rec spe f1 geo iba sup**  1.000 0.943 0.813 0.971 0.876 0.777 71079 0.024 0.813 0.943 0.047 0.876 0.757 123  0.998 0.943 0.813 0.969 0.876 0.777 71202 |

|  |
| --- |
| **SMOTE - RandomForestClassifier** |
| **pre rec spe f1 geo iba sup**  1.000 1.000 0.780 1.000 0.883 0.797 71079 0.889 0.780 1.000 0.831 0.883 0.763 123  0.999 0.999 0.781 0.999 0.883 0.797 71202 |

|  |
| --- |
| **SMOTE - LinearSVC** |
| **pre rec spe f1 geo iba sup**  1.000 0.979 0.902 0.989 0.940 0.890 71079 0.068 0.902 0.979 0.126 0.940 0.876 123  0.998 0.978 0.903 0.988 0.940 0.890 71202 |

|  |
| --- |
| **SMOTE - GaussianNB** |
| **pre rec spe f1 geo iba sup**  1.000 0.975 0.837 0.987 0.904 0.828 71079 0.055 0.837 0.975 0.104 0.904 0.805 123  0.998 0.975 0.838 0.986 0.904 0.828 71202 |

|  |
| --- |
| **BalancedBaggingClassifier - RandomForestClassifier** |
| **pre rec spe f1 geo iba sup**  1.000 0.987 0.870 0.993 0.926 0.868 71079 0.101 0.870 0.987 0.180 0.926 0.848 123  0.998 0.986 0.870 0.992 0.926 0.868 71202 |

|  |
| --- |
| **BalancedBaggingClassifier - LinearSVC** |
| **pre rec spe f1 geo iba sup**  1.000 0.976 0.894 0.988 0.934 0.880 71079 0.060 0.894 0.976 0.112 0.934 0.865 123  0.998 0.976 0.894 0.986 0.934 0.880 71202 |

|  |
| --- |
| **BalancedBaggingClassifier - GaussianNB** |
| **pre rec spe f1 geo iba sup**  1.000 0.967 0.846 0.983 0.904 0.828 71079 0.043 0.846 0.967 0.082 0.904 0.808 123  0.998 0.967 0.846 0.982 0.904 0.828 71202 |

Τα συμπεράσματα που μπορεί να εξαχθούν είναι τα εξής:

* Η χρήση της μεθόδου NearMiss παρουσιάζει γενικά φτωχά αποτελέσματα, με εξαίρεση όταν γίνεται χρήση Naïve Bayes που δίνει geometric mean > 0.85.
* Η χρήση της μεθόδου SMOTE αν και αυξάνει αρκετά τον χρόνο εκπαίδευσης δίνει καλά αποτελέσματα με σχετικά μικρή διακύμανση και geometric mean > 0.85 για όλους τους ταξινομητές. Επίσης παρουσιάζει τα καλύτερα score για τις μετρικές geometric mean και index balanced accuracy όταν εφαρμόζεται ο LinearSVC.
* Η χρήση της μεθόδου EasyEnsemble είναι πολύ γρήγορη σε χρόνο εκπαίδευσης και δίνει καλά αποτελέσματα με μικρή διακύμανση. Οι τιμές της μετρικής geometric mean είναι μεγαλύτερες του 0.9, ενώ και αυτές για την μετρική index balanced accuracy είναι αρκετά υψηλές για τις δύο κλάσεις.