ΝΕΑ ΜΕΘΟΔΟΛΟΓΊΑ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΕΠΙΒΛΕΠΟΜΕΝΗΣ ΚΑΙ ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗΣ ΕΚΜΑΘΗΣΗΣ

Διπλωματική Εργασία

του

Μιχαλόπουλου Μάριου

26 Φεβ 2020

Περίγραμμα παρουσίασης

- Σκοπός της εργασίας
- Περιγραφή αλγορίθμων
- Περιγραφή προτεινόμενης μεθοδολογίας
- Περιγραφή συνόλων δεδομένων και αποτελέσματα
- Συμπεράσματα

Σκοπός της εργασίας

- Κατασκευή μοντέλων πρόβλεψης βασισμένα σε τέσσερα διαφορετικά σύνολα δεδομένων με την χρήση διάφορων αλγορίθμων κατηγοριοποίησης και παλινδρόμησης.

- Σύγκριση των παραπάνω μοντέλων ταξινόμησης, με μια προτεινόμενη μεθοδολογία που συνδυάζει τον αλγόριθμο Κ-μέσων με τα αντίστοιχα μοντέλα και με μια μέθοδο επιλογής χαρακτηριστικών.

Αλγόριθμοι που θα χρησιμοποιηθούν

1. Νευρωνικά δίκτυα (Multilayer Perceptron)
2. Δέντρα απόφασης (Decision Trees)
3. Λογιστική παλινδρόμηση (Logistic Regression)
4. Γραμμική παλινδρόμηση (Linear Regression)
5. Αλγόριθμος Κ-μέσων (K-Means)

Γραμμική παλινδρόμηση

- Απλή Γραμμική παλινδρόμηση

$$y=\beta_0+\beta_1 x+\varepsilon$$

β_0 το σημείο τομής της ευθείας με τον άξονα, β_1 η κλίση της ευθείας και ε μια τυχαία μεταβλητή λάθους

- Πολλαπλή Γραμμική παλινδρόμηση

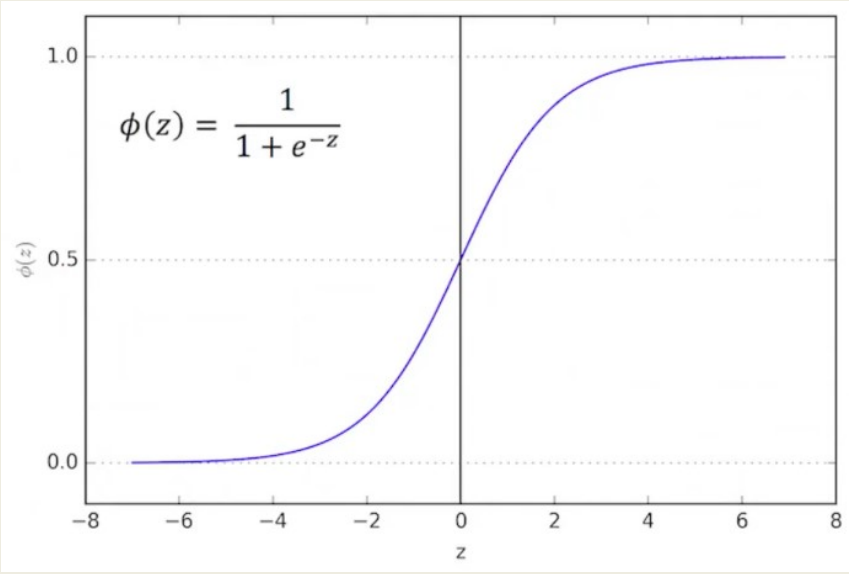
$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k+\varepsilon$$

- Τα λάθη θεωρούμε ότι έχουν μέσο όρο 0 και διακύμανση σ^2 που είναι άγνωστη.
- Επιπλέον, θεωρούμε ότι τα λάθη δεν σχετίζονται μεταξύ τους

Λογιστική παλινδρόμηση

- γραμμικό μοντέλο για ταξινόμηση και όχι για παλινδρόμηση
- μεταβλητή *y* συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δύο τιμές 0 ή 1)

$$z=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k \Longrightarrow \Phi(z)$$



Δέντρο απόφασης

- Οικοδομεί ένα δέντρο από ένα σύνολο δεδομένων χρησιμοποιώντας την έννοια του κέρδους της πληροφορίας.

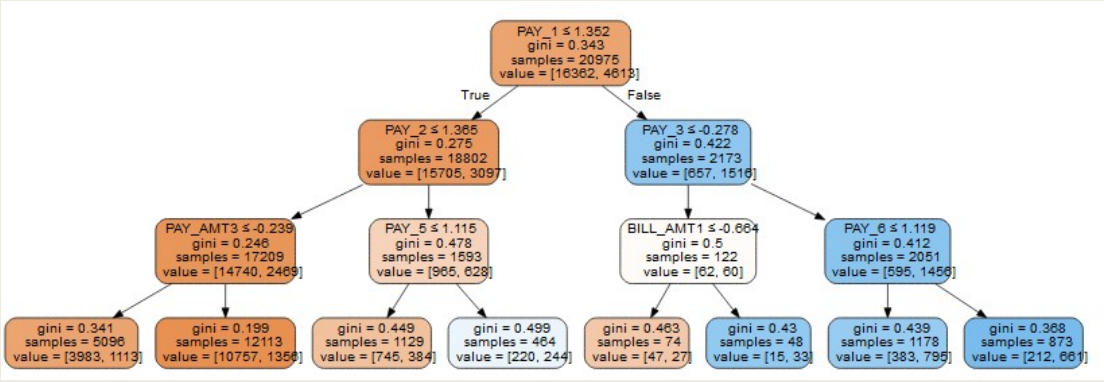
Πιο γνωστοί αλγόριθμοι:

- ID3 (Quinlan J.R., 1986) και C4.5 (Quinlan J. R., 2014)

Αντίστοιχος αλγόριθμος που θα χρησιμοποιηθεί:

- Δέντρα Ταξινόμησης και παλινδρόμησης (C.A.R.T)

Δέντρο απόφασης



C.A.R.T: Χρήση δείκτη Gini ή εντροπία για το διαχωρισμό

➢ <u>Εντροπία</u>

Εντροπία είναι το μέτρο της τυχαιότητας της πληροφορίας. Όσο μεγαλύτερη, τόσο δυσκολότερο να εξάγουμε συμπεράσματα από την πληροφορία.

➢ <u>Δείκτης Gini</u>

$\text{Gini}(D)=1-\sum_{i=1}^{m} Pi^2$ με P_i η πιθανότητα η πλειάδα D να ανήκει στην κλάση C_i

Διαχωρίζουμε τα δεδομένα μας έτσι ώστε να ελαχιστοποιείται

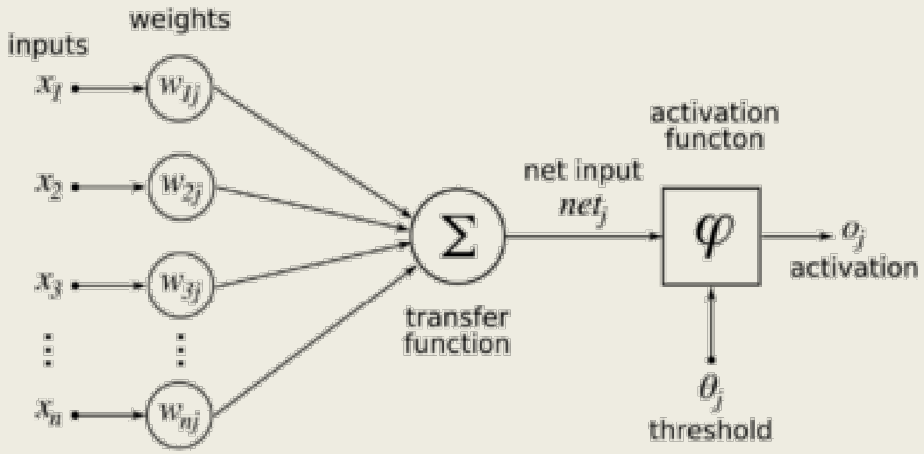
$$\text{Gini}_A(D)=\frac{|D1|}{|D|}\text{Gini}(D_1)+\frac{|D2|}{|D|}\text{Gini}(D_2)$$

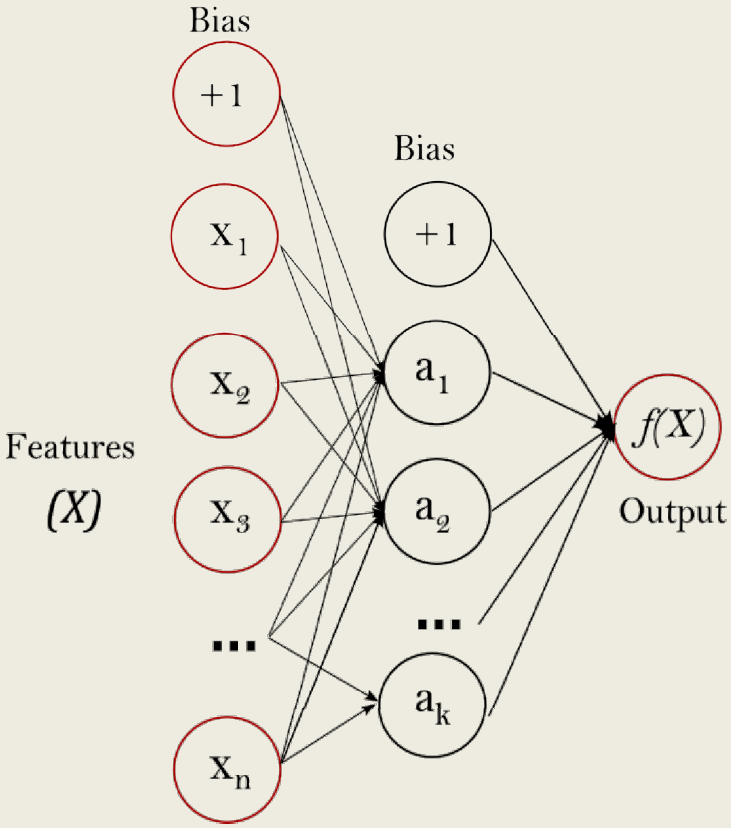
Νευρωνικό Δίκτυο

- Δίκτυο από «νευρώνες»
- Αποτελείται από τριών ειδών επίπεδα

(*input layer, hidden layer , output layer*)

- *Input layer*: κάθε κόμβος αντιστοιχεί σε ένα χαρακτηριστικό των δεδομένων
- *Hidden layer*: κάθε κόμβος είναι το σταθμισμένο άθροισμα των συνδεδεμένων κόμβων, επαυξημένο κατά μια σταθερά (bias).
- Η τιμή που εκπέμπει κάθε κόμβος περνάει από συνάρτηση ενεργοποίησης

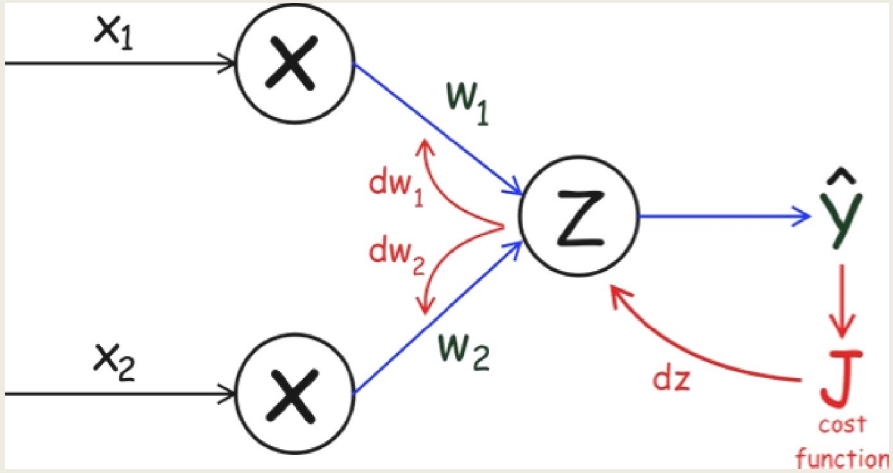


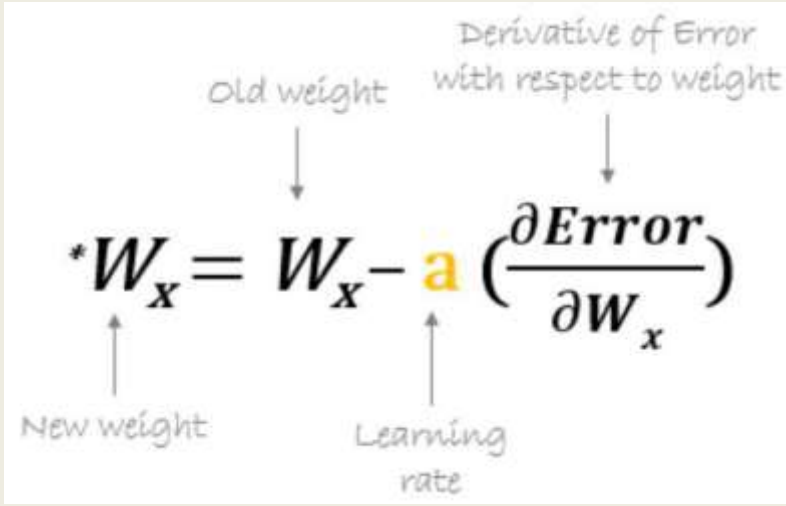


Νευρωνικό Δίκτυο

Πως μαθαίνει το δίκτυο;

- Χρήση backpropagation αλγορίθμου για διάδοση του λάθους
- Όταν μια καταγραφή έχει κατηγοριοποιηθεί λάθος, υπολογίζεται το σφάλμα.
- Ανανεώνονται αναδρομικά τα βάρη όλων των κόμβων που οδήγησαν σε αυτή την απόφαση.





Αλγόριθμος Κ-μέσων

- Ανήκει στους αλγόριθμους μη εποπτευόμενης μάθησης
- Διαχωρίζει τα δείγματα σε ομάδες (συμπλέγματα) Κ ίσων διακυμάνσεων
- έχει στόχο να επιλέξει τα κεντροειδή που ελαχιστοποιούν ένα κριτήριο αδράνειας (το σύνολο των τετραγώνων μεταξύ των ομάδων)

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (|x_i - \mu_j|^2)$$

Προτεινόμενη Μεθοδολογία

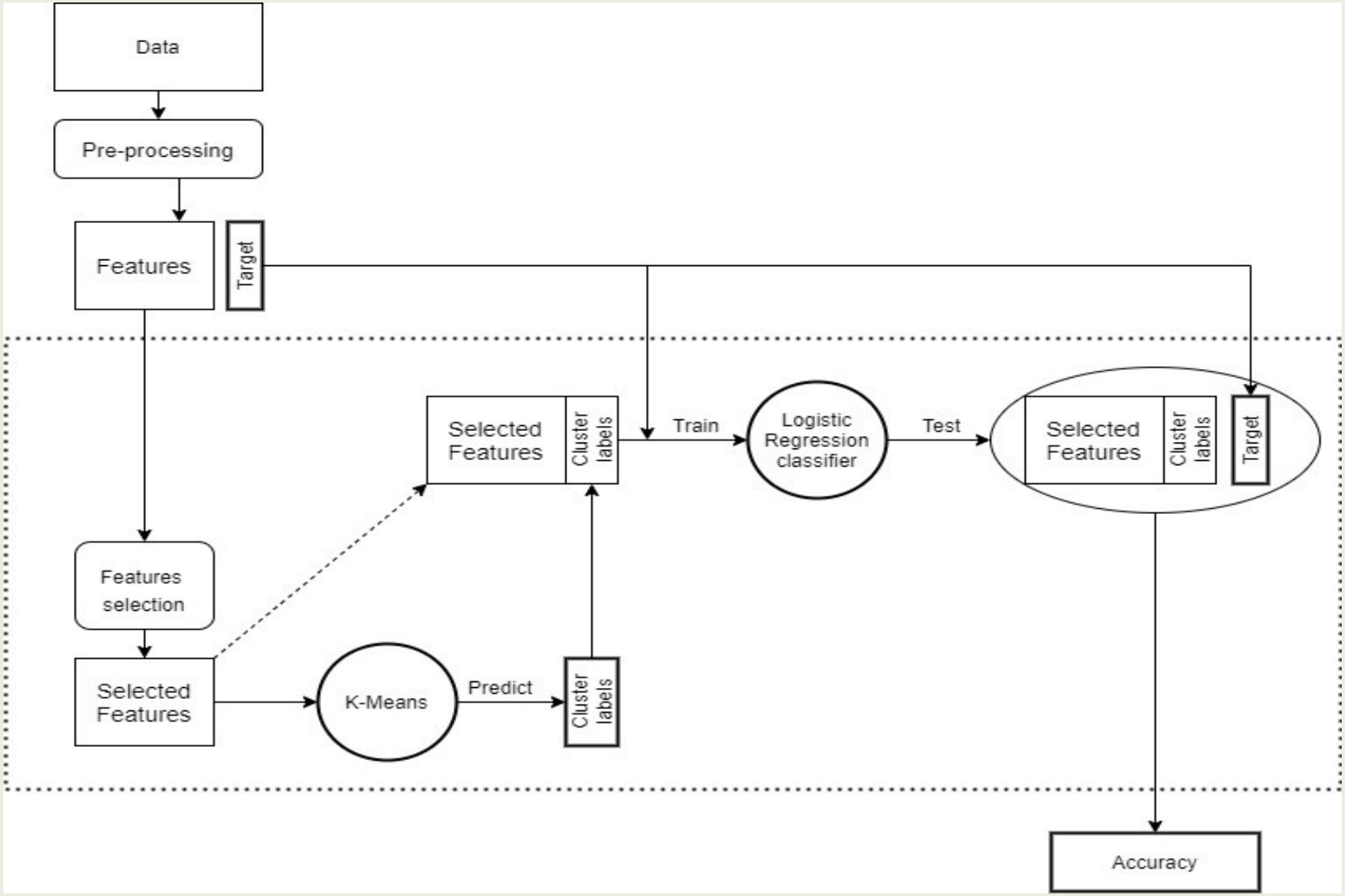
Χρήση τριών αλγορίθμων:

- Επιλογή χαρακτηριστικών
- Αλγόριθμος Κ-μέσων
- Ένας αλγόριθμός ταξινόμησης (Λογιστική παλινδρόμηση)

Στοιχεία μεθοδολογίας

- Επιλογή πλήθους συστάδων με βάση τις τιμές του χαρακτηριστικού στόχος
- Αναζήτηση Ν καλύτερων χαρακτηριστικών με την χρήση μεθόδου ANOVA
- Υλοποίηση «πρώιμου σταματήματος»

Προτεινόμενη Μεθοδολογία

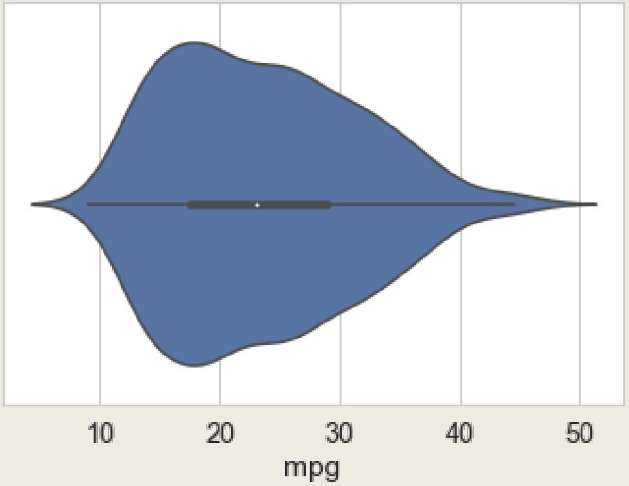


Σύνολο δεδομένων Auto MPG

Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	517160

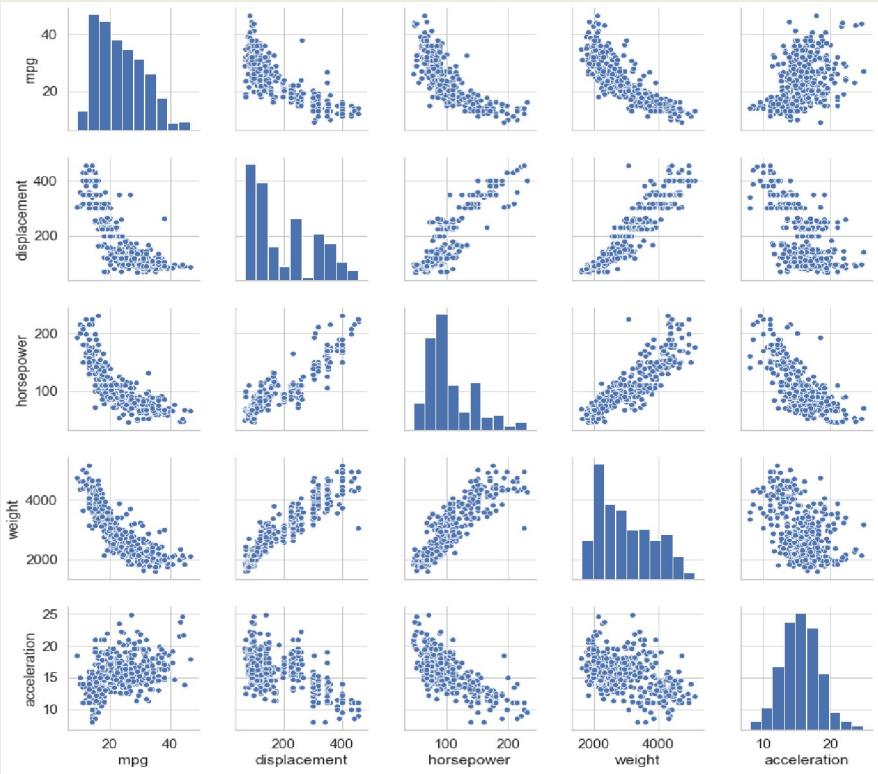
Το σύνολο δεδομένων Auto MPG αφορά την αυτονομία καυσίμου για αυτοκίνητα που κινούνται εντός πόλης. Περιέχει 398 εγγραφές οχημάτων.

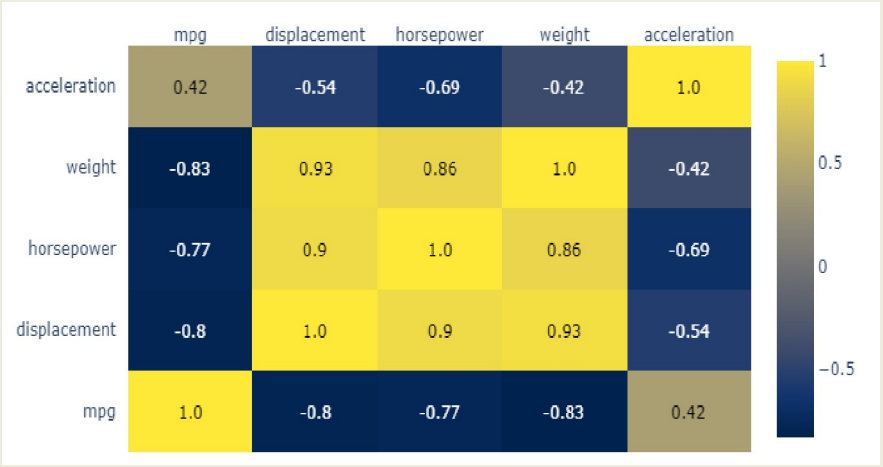
➢Χαρακτηριστικά: *mpg, displacement, horsepower, weight, acceleration, origin, model year, car name*



➢Πεδία χωρίς τιμές:
6 πεδία στο *horsepower* → συμπλήρωση με μέσο όρο
8 πεδία στο *mpg* → αφαίρεση

Σύνολο δεδομένων Auto MPG





- Υψηλή γραμμική συσχέτιση μεταξύ *weight, displacement, horsepower* με *mpg*

- Υψηλή γραμμική συσχέτιση *weight, displacement, horsepower* μεταξύ τους

<u>Μετασχηματισμός δεδομένων</u>

- Αφαίρεση *displacement, horsepower*

- Κωδικοποίηση σε 0 και 1 χαρακτηριστικού *origin*

- Κανονικοποίηση σε μέσο όρο 0, τυπ. απόκλιση 1

- Υποσύνολο εκπαίδευσης – υποσύνολο δοκιμής 70% - 30%

Σύνολο δεδομένων Auto MPG

Αποτελέσματα

- Υποσύνολο δοκιμής – 10 Fold CV

Models	MAE	MSE	RMSE	CV MAE	CV MSE	CV RMSE
Linear Regression	2.533	11.211	3.348	2.748	12.804	3.578
Decision Tree	2.407	12.784	3.576	2.733	14.801	3.847
Neural Network	1.924	6.794	2.607	2.451	11.271	3.357

- Ανάλυση της διακύμανσης

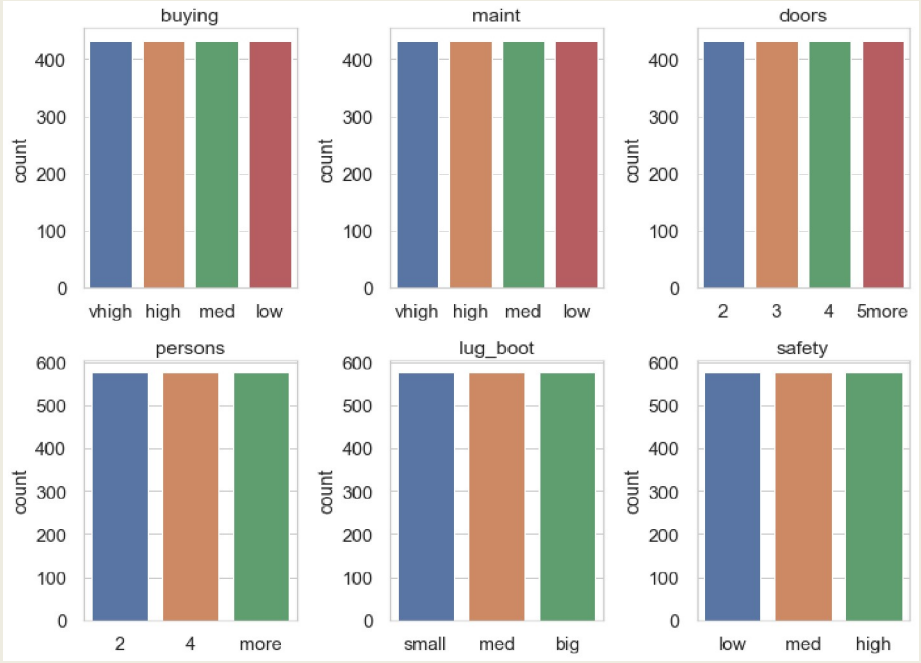
Models	Linear Regression	Decision Tree	Neural Network
Linear Regression	-	0.017	0.706
Decision Tree	0.017	-	0.648
Neural Network	0.706	0.648	-

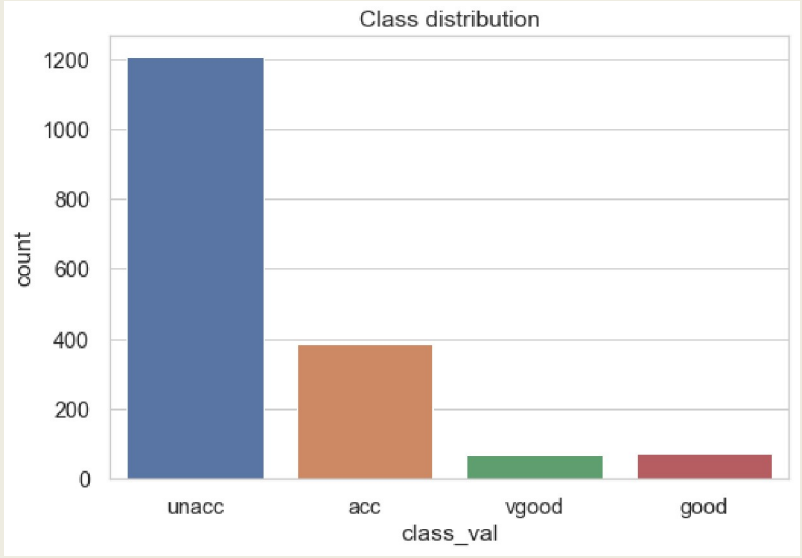
Σύνολο δεδομένων Car Evaluation

Data Set Characteristics:	Multivariate	**Number of Instances:**	1728	**Area:**	N/A
Attribute Characteristics:	Categorical	**Number of Attributes:**	6	**Date Donated**	1997-06-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1126810

Το μοντέλο αυτό κατατάσσει τα αυτοκίνητα σε 4 κατηγορίες, με βάση έξι ποιοτικά χαρακτηριστικά τους. Περιέχει 1728 εγγραφές.

➢<u>Χαρακτηριστικά:</u> *buying, maint, doors, persons, lug_boot, safety, class_val*



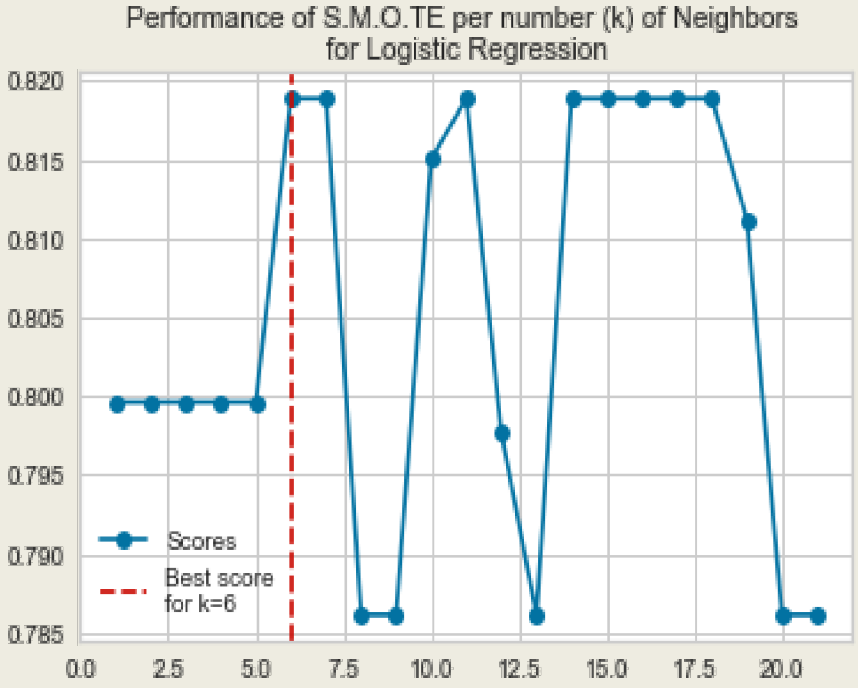


70% (1210) *unacc*, 22,2% (384) *acc*, 3,99% (69) *good*, 3,76% (65) *vgood*

Σύνολο δεδομένων Car Evaluation

Μετασχηματισμός δεδομένων

- Κωδικοποίηση όλων των χαρακτηριστικών πρόβλεψης σε 0 και 1.
- Νέο πλήθος χαρακτηριστικών: 21
- Κωδικοποίηση χαρακτηριστικού στόχος με τις τιμές από 0 έως 3, για κάθε μια από τις τέσσερις κατηγορίες
- Υποσύνολο εκπαίδευσης – υποσύνολο δοκιμής 70% - 30%



Μη ισορροπημένες κλάσεις

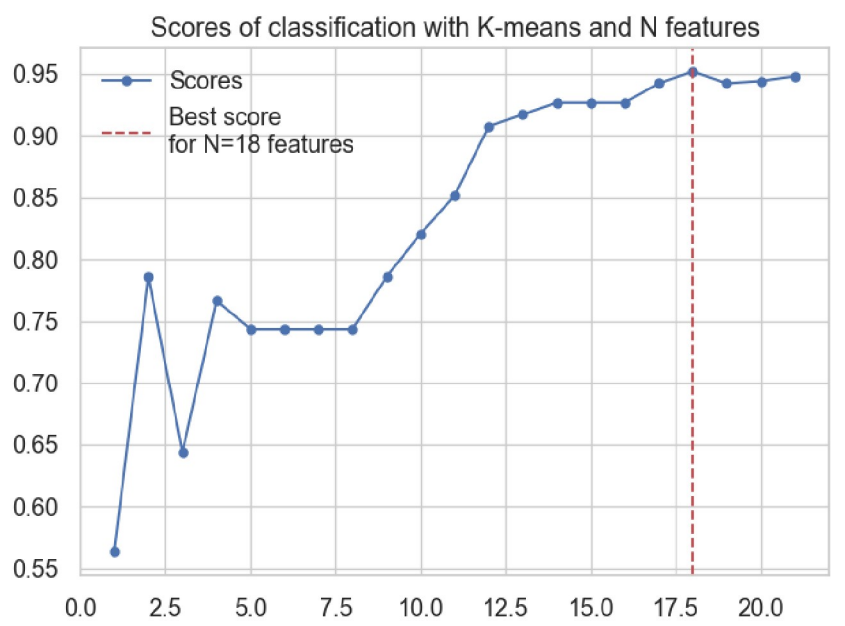
•Χρήση SMOTE για τις 3 κατηγορίες με τις λιγότερες εγγραφές

•Χρήση λογιστικής παλινδρόμησης κατά την διερεύνηση

• Καλύτερα αποτελέσματα για K=6

Σύνολο δεδομένων Car Evaluation

Αποτελέσματα



Models	Accuracy	Precision	Recall	F_1	AUC
Logistic Regression	0.792	0.381	0.455	0.397	0.949
Logistic Regression + Kmeans	0.931	0.802	0.916	0.849	0.990
Decision Tree	0.715	0.530	0.679	0.528	0.924
Decision Tree + Kmeans	0.805	0.615	0.803	0.671	0.936
Neural Network	0.898	0.790	0.803	0.796	0.983
Neural Network + Kmeans	0.960	0.879	0.934	0.904	0.996

Models	CV Accuracy	CV Precision	CV Recall	CV F_1	CV AUC
Logistic Regression	0.833	0.684	0.680	0.653	0.974
Logistic Regression + Kmeans	0.844	0.771	0.740	0.714	0.977
Decision Tree	0.760	0.434	0.475	0.426	0.861
Decision Tree + Kmeans	0.790	0.564	0.558	0.524	0.921
Neural Network	0.894	0.877	0.882	0.861	0.977
Neural Network + Kmeans	0.873	0.816	0.834	0.797	0.966

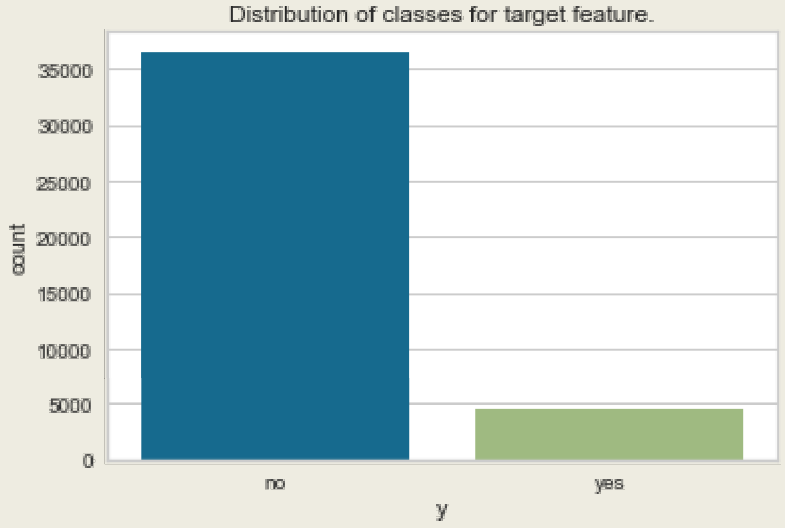
- Χρήση της μετρικής AUC για αξιολόγηση των μοντέλων
- Αύξηση όλων των μετρικών για όλα τα «σύνθετα» μοντέλα στο υποσύνολο δοκιμής
- NN χειρότερη απόδοση στο 10 fold CV με K-means

Σύνολο δεδομένων Bank Marketing

Data Set Characteristics:	Multivariate	**Number of Instances:**	45211	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	17	**Date Donated**	2012-02-14
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	1116654

Περιλαμβάνει πραγματικά δεδομένα που συλλέχθηκαν από πορτογαλική τράπεζα, από το Μάιο του 2008 έως τον Νοέμβριο του 2010, με συνολικά 41188 τηλεφωνικές επαφές.

➢Χαρακτηριστικά: *age, job, marital, education, default, housing, loan, contact, month, day_of_week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, y*



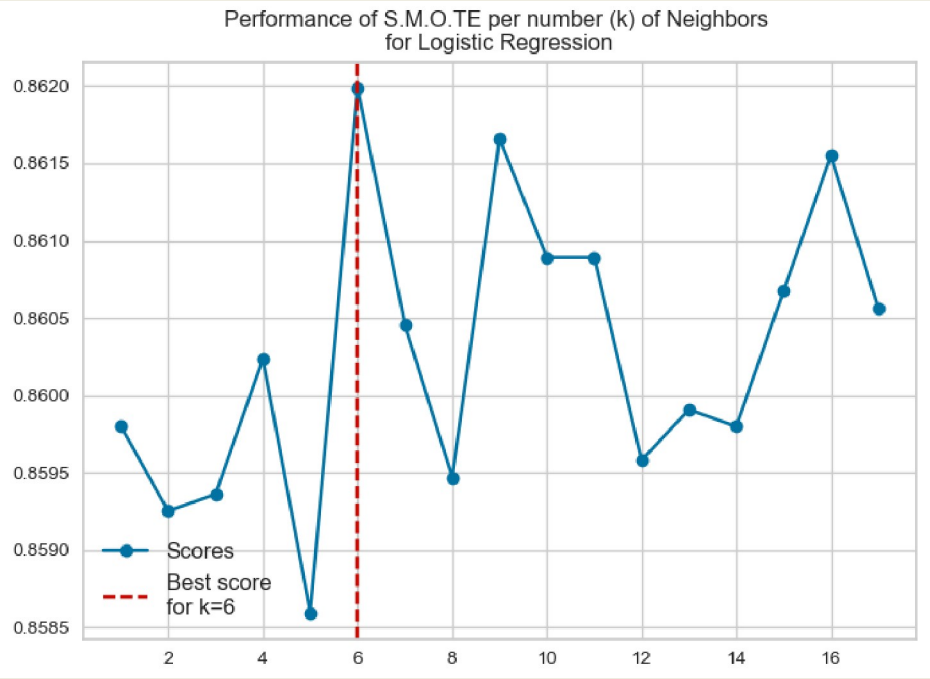
Yes: 4640 (11,26%) } 3858 (12,65%)
No: 36548 (88,74%) } 26620 (87,34%)

12 διπλοεγγραφές ⟹ αφαίρεση
τιμές unknown ⟹ αφαίρεση
(*job, marital, education, default, housing, loan*)

Σύνολο δεδομένων Bank Marketing

Μετασχηματισμός δεδομένων

- Αφαίρεση χαρακτηριστικού *nr.employed* λόγω υψηλής συσχέτισης με το *emp.var.rate*
- Κωδικοποίηση με 0 και 1 για τα κατηγορικά χαρακτηριστικά
- Συνολικά προκύπτουν 81 χαρακτηριστικά
- Κανονικοποίηση σε μέσο όρο 0 και σε μοναδιαία τυπική απόκλιση
- Υποσύνολο εκπαίδευσης – υποσύνολο δοκιμής 70% - 30%

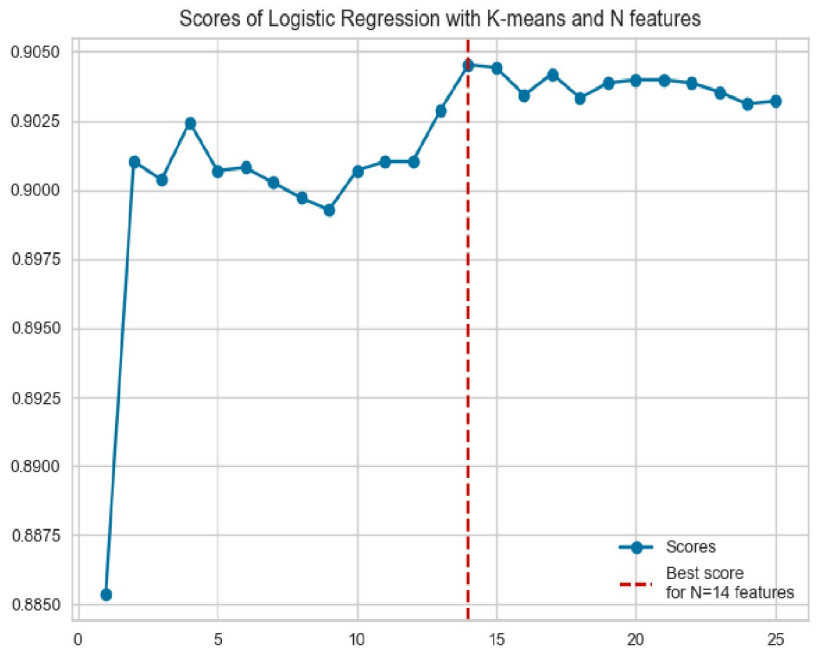


Μη ισορροπημένες κλάσεις

- Χρήση SMOTE για την κλάση yes
- Χρήση λογιστικής παλινδρόμησης κατά την διερεύνηση
- Καλύτερα αποτελέσματα για K=6

Σύνολο δεδομένων Bank Marketing

Αποτελέσματα



Models	Accuracy	Precision	Recall	F_1	AUC
Logistic Regression	0.862	0.464	0.885	0.609	0.937
Logistic Regression + Kmeans	0.855	0.450	0.854	0.589	0.931
Decision Tree	0.837	0.411	0.805	0.544	0.889
Decision Tree + Kmeans	0.840	0.419	0.827	0.556	0.881
Neural Network	0.887	0.527	0.718	0.608	0.928
Neural Network + Kmeans	0.846	0.434	0.899	0.586	0.936

Models	CV Accuracy	CV Precision	CV Recall	CV F_1	CV AUC
Logistic Regression	0.822	0.636	0.303	0.327	0.862
Logistic Regression + Kmeans	0.849	0.663	0.341	0.362	0.913
Decision Tree	0.775	0.225	0.295	0.190	0.630
Decision Tree + Kmeans	0.781	0.359	0.340	0.258	0.658
Neural Network	0.764	0.365	0.252	0.203	0.823
Neural Network + Kmeans	0.788	0.363	0.260	0.209	0.849

- Χρήση της μετρικής AUC για αξιολόγηση των μοντέλων
- Αύξηση AUC κατά το 10-fold CV με την χρήση του αλγορίθμου K-μέσων
- Αναζήτηση καλύτερων υπερ-παραμέτρων για Δέντρο απόφασης και ενδεχομένως για Νευρωνικό Δίκτυο

Σύνολο δεδομένων Bank Marketing

Αποτελέσματα

- Χωρίς τις unknown τιμές

Models	CV Accuracy	CV Precision	CV Recall	CV F_1	CV AUC
Logistic Regression	0.822	0.636	0.303	0.327	0.862
Logistic Regression + Kmeans	0.849	0.663	0.341	0.362	0.913
Decision Tree	0.775	0.225	0.295	0.190	0.630
Decision Tree + Kmeans	0.781	0.359	0.340	0.258	0.658
Neural Network	0.764	0.365	0.252	0.203	0.823
Neural Network + Kmeans	0.788	0.363	0.260	0.209	0.849

- Με τις unknown τιμές

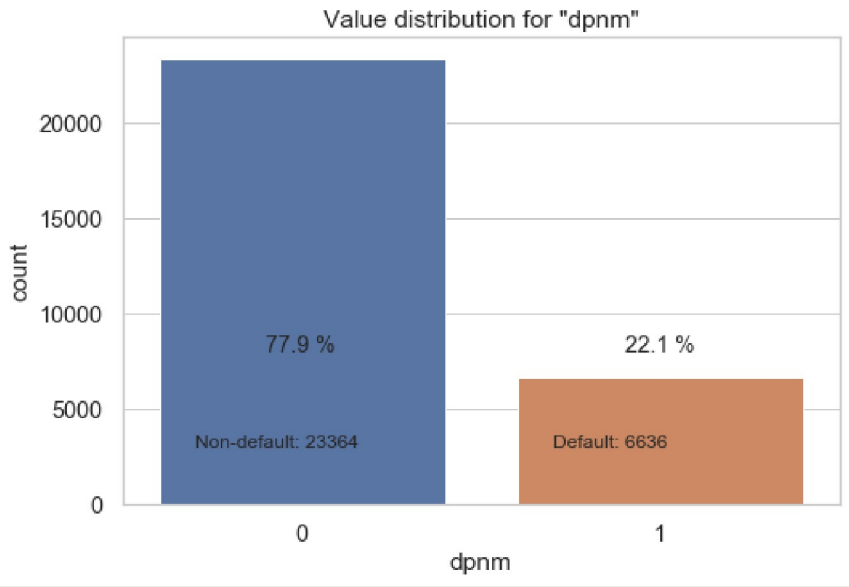
Models	CV Accuracy	CV Precision	CV Recall	CV F_1	CV AUC
Logistic Regression	0.838	0.604	0.305	0.316	0.860
Logistic Regression +Kmeans	0.869	0.658	0.343	0.365	0.925
Decision Tree	0.771	0.317	0.271	0.225	0.643
Decision Tree + Kmeans	0.748	0.393	0.317	0.281	0.614
Neural Network	0.770	0.244	0.199	0.129	0.797
Neural Network + Kmeans	0.813	0.339	0.253	0.205	0.873

Σύνολο δεδομένων
Default of credit card clients

Data Set Characteristics:	Multivariate	**Number of Instances:**	30000	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	24	**Date Donated**	2016-01-26
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	439264

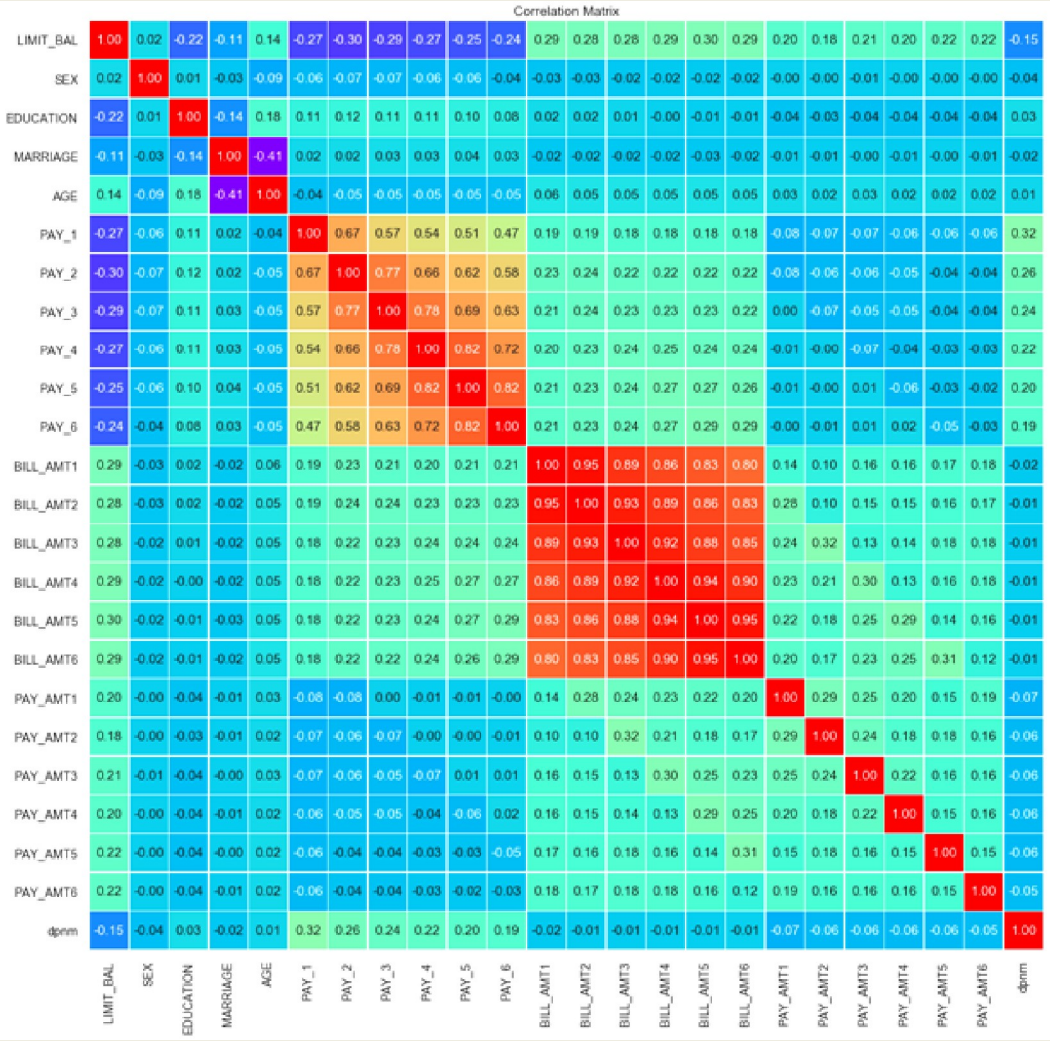
Περιέχει δεδομένα πελατών ταϊβανέζικων τραπεζών, οι οποίοι είχαν στην κατοχή τους και χρησιμοποιούσαν πιστωτικές κάρτες. Περιέχει 30000 εγγραφές

➢Χαρακτηριστικά: *limit_bal, sex, education, marriage, age, pay_1, pay_2, pay_3, pay_4, pay_5, pay_6, bill_amt1, bill_amt2, bill_amt3, bill_amt4, bill_amt5, bill_amt6, pay_amt1, pay_amt2, pay_amt3, pay_amt4, pay_amt5, pay_amt6, dpnm*



Default: 6636 (22,1%)
Non Default: 23364 (77,9%)

Σύνολο δεδομένων
Default of credit card clients



Μετασχηματισμός δεδομένων

• Αφαίρεση *BILL_AMT2* έως *BILL_AMT6* λόγω υψηλής συσχέτισης με το *BILL_AMT1*

•Μοναδικές τιμές:
MARRIAGE: 4→3, *EDUCATION:* 6→4

• Κωδικοποίηση σε 0 και 1 των *EDUCATION, MARRIAGE, SEX, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6*

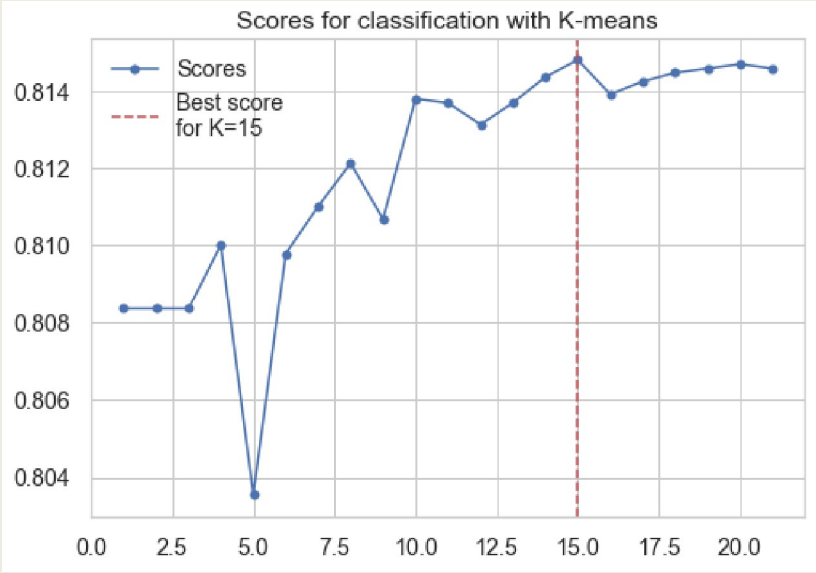
• Το πλήθος των χαρακτηριστικών αυξήθηκε στα 82

• Κανονικοποίηση σε μηδενικό μέσο όρο και μοναδιαία τυπική απόκλιση.

• Υποσύνολο εκπαίδευσης – υποσύνολο δοκιμής 70% - 30%

Σύνολο δεδομένων
Default of credit card clients

Αποτελέσματα



Models	Accuracy	Precision	Recall	F_1	AUC
Logistic Regression	0.817	0.676	0.353	0.463	0.77
Logistic Regression + Kmeans	0.815	0.672	0.341	0.452	0.77
Decision Tree	0.814	0.669	0.337	0.448	0.74
Decision Tree + Kmeans	0.815	0.674	0.340	0.452	0.72
Neural Network	0.814	0.650	0.371	0.472	0.77
Neural Network + Kmeans	0.817	0.668	0.371	0.477	0.76

Models	CV Accuracy	CV Precision	CV Recall	CV F_1	CV AUC
Logistic Regression	0.779	0	0	0	0.639
Logistic Regression+Kmeans	0.818	0.674	0.345	0.456	0.757
Decision Tree	0.818	0.681	0.334	0.447	0.748
Decision Tree + Kmeans	0.820	0.679	0.354	0.464	0.723
Neural Network	0.761	0.546	0.204	0.226	0.657
Neural Network + Kmeans	0.819	0.674	0.355	0.464	0.763

- Υποσύνολο δοκιμής – 10 fold CV πολύ κοντινά αποτελέσματα στην ευστοχία
- Ο αλγόριθμος Κ-μέσων βοήθησε σημαντικά κατά το 10 fold CV
- Όχι μεγάλες διαφορές στα αποτελέσματα των μοντέλων

Συμπεράσματα

Models	Data sets			Average
	Car Evaluation	Bank Marketing	Default of credit card clients	
Logistic Regression	0.833	0.822	0.779	0.811
Logistic Regression + Kmeans	0.844	0.849	0.818	0.837
Decision tree	0.760	0.775	0.818	0.784
Decision tree + Kmeans	0.790	0.781	0.820	0.797
Neural Network	0.894	0.764	0.761	0.806
Neural Network + Kmeans	0.873	0.788	0.819	0.826

<u>Πλεονεκτήματα</u>

- Γενικά καλύτερα αποτελέσματα σε σχέση με τα απλά μοντέλα
- Δεν απαιτεί ιδιαίτερες ρυθμίσεις κατά την χρήση της
- Αποφυγή προεπιλογής πλήθους συστάδων, μόνο επιλογή χαρακτηριστικών
- Είναι αρκετά γρήγορη

<u>Μειονεκτήματα</u>

- Ο K-means δεν είναι πάντα ο πιο κατάλληλος αλγόριθμος για κάθε σύνολο δεδομένων.
- Εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά που θα επιλεχτούν
- Διερεύνηση αν για κάποιο άλλο αριθμό κεντροειδών πετυχαίνεται καλύτερο αποτέλεσμα

Ευχαριστώ για την προσοχή σας!