ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ

ΑΝΑΣΤΑΣΙΟΥ ΘΕΟΔΩΡΑ itp19103
ΓΕΩΡΓΙΑΔΗΣ ΜΑΡΙΟΣ itp19105
ΛΑΓΟΥ ΟΥΡΑΝΙΑ itp19122

Εργασία στο μάθημα **Εξόρυξης Δεδομένων και Επιχειρηματική Ευφυΐα**

Περιεχόμενα

Γενικά

Η εργασία υλοποιήθηκε σε περιβάλλον Visual Studio Code , σε γλώσσα Python 3.
Έχουν δημιουργηθεί δύο πηγαία αρχεία κώδικα το ένα για την προεπεξεργασία των test δεδομένων και ένα για την προεπεξεργασία των training δεδομένων, της κατηγοριοποίησης και της συσταδοποίησης.
Στον φάκελο περιέχονται 2 πηγαία αρχεία τα οποία εκτελούνται με τις εξής εντολές :

1) python3 eksoriksi1.py
(περιέχει την προεπεξεργασία - οπτικοποίηση - εκπαίδευση - κατηγοριοποίηση - συσταδοποίηση των δεδομένων εκπαίδευσης και προβλέπει και αποθηκεύει τα τελικά αποτελέσματα)

2) python3 test_file.py
(περιέχει την προεπεξεργασία και την δημιουργία του αρχείου των δεδομένων ελέγχου)

Στον φάκελο περιέχεται επίσης

- το αρχείο Report.csv στο οποίο καταγράφονται στατιστικά για τους αλγόριθμους κατηγοριοποίησης.
- το αρχείο hello.csv με τα δεδομένα εκπαίδευσης
- το αρχείο hellotest.csv με τα δεδομένα ελέγχου
- το αρχείο ac_statetime_test.csv με τις τελικές προβλέψεις

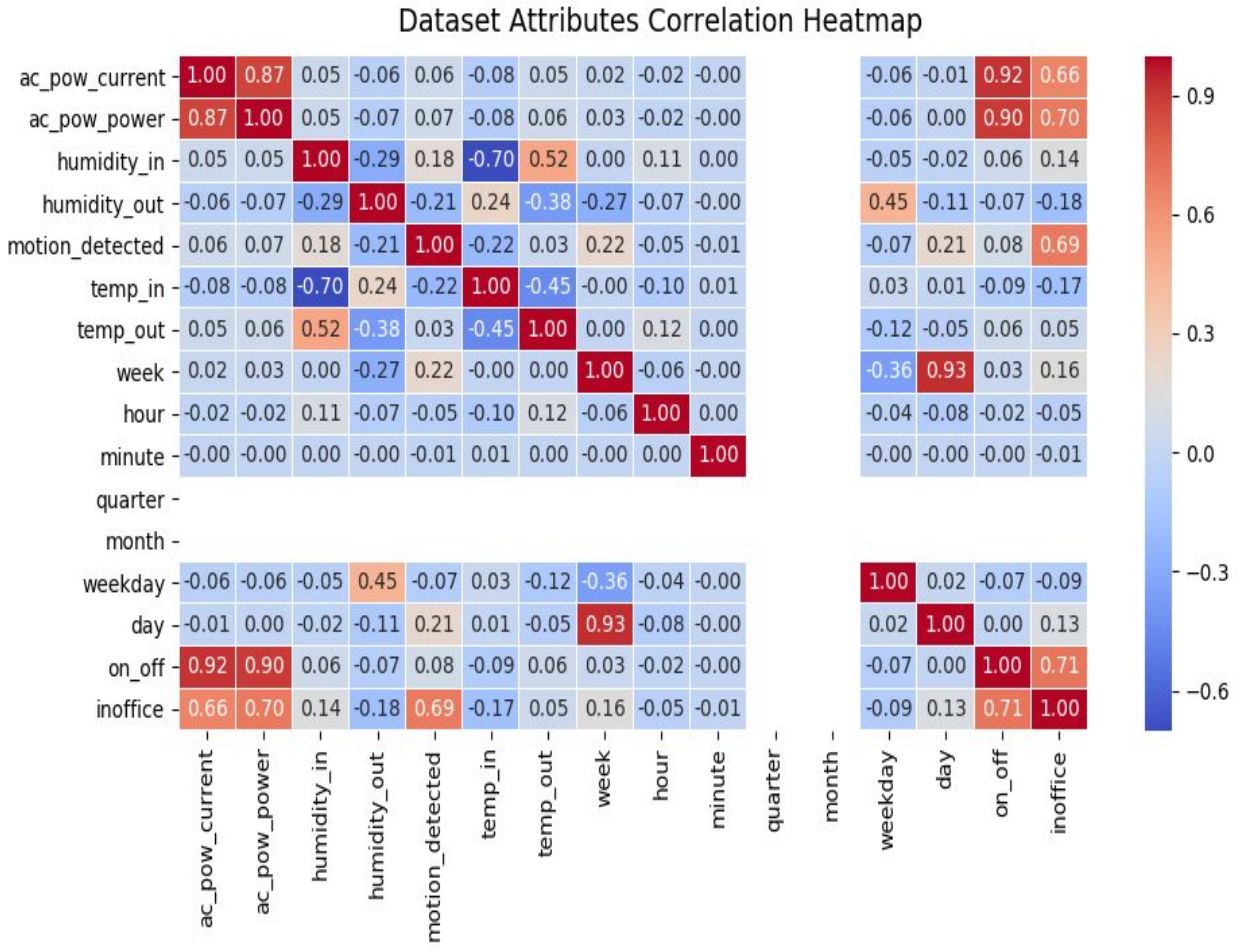
A)Preparation/Προετοιμασία Δεδομένων

Μετά από μελέτη και κατανόηση των δεδομένων που περιέχονται στα αρχικώς δοθέντα ξεχωριστά αρχεία, τα μετατρέψαμε σε ένα ενιαίο αρχείο. Το ενιαίο πλέον αρχείο, που προκύπτει αρχικά με πολλές κενές θέσεις, περιέχει συνολικά 8 στήλες, τις εξής 7 που αφορούν τα δεδομένα “ac_pow_power”, ”ac_pow_current”, ”humidity_in”, ”humidity_out”, ”temp_in”, ”temp_out”, ”motion_detected”, τα οποία λαμβάνονται απο τα ξεχωριστά αρχεία, καθώς και την κοινή στήλη σε όλα τα αρχεία “statetime”. Στην συνέχεια φροντίζουμε να υπάρχει μια τιμή για κάθε αισθητήρα για κάθε χρονική περίοδο όπως μας ζητήθηκε, ομαδοποιώντας το αρχείο σε χρονικά διαστήματα ενδεικτικά του 1 λεπτού, ενώ ταυτόχρονα υπολογίζουμε τον μέσο όρο για αυτό το διάστημα, εφόσον υπάρχουν διαθέσιμες τιμές για αυτό, αντικαθιστώντας έτσι τις πολλές τιμές που είχαμε σε κάθε διάστημα με μια μέση τιμή. Για την διαδικασία αυτή χρησιμοποιήσαμε εντολή groupby σε συνδυασμό με συνάρτηση Grouper.

Μετά την ομαδοποίηση των στηλών με βάση τον χρόνο, συνεχίζουμε με το καθάρισμα και το συμπλήρωμα των δεδομένων. Για την στήλη motion_detected αντικαθιστούμε την τιμή “ 'EA674E” με 1 θεωρώντας ότι όταν είναι 1 καταγράφηκε κίνηση και όταν είναι nan 0, δηλαδή ότι δεν καταγράφηκε κάποια κίνηση. Παρατηρήσαμε ότι στις στήλες “ac_pow_power”, ”ac_pow_current”, υπήρχαν τιμές “unknown” για τις οποίες θεωρούμε ότι ίσως έπαθε κάποια βλάβη ο συγκεκριμένος αισθητήρας και τις αντικαθιστούμε με 0. Επιπρόσθετα παρατηρούμε ότι σε ένα μεγάλο ποσοστό των δεδομένων οι στήλες “ac_pow_power”, ”ac_pow_current” ήταν κενές(nan), έτσι θεωρήσαμε ότι έπεφταν σε ημερομηνίες που το κλιματιστικό ήταν κλειστό, π.χ σαββατοκύριακα είτε είχαν κάποια βλάβη, οπότε θεωρώντας το κλιματιστικό κλειστό, αντικαθιστούμε τις κενές τιμές των δύο στηλών με 0. Ακολούθως θεωρούμε ότι η στήλη “ac_pow_power” μας δίνει την πληροφορία εάν το κλιματιστικό είναι on/off έτσι χρησιμοποιώντας αυτές τις δύο στήλες δημιουργήσαμε μια καινούρια “on_off” στην οποία άν οι τιμές ac_pow_power+ac_pow_current > 0 τότε το κλιματιστικό είναι ανοιχτό(1) αλλιώς το κλιματιστικό είναι κλειστό(0). Στην συνέχεια δημιουργήσαμε τις στήλες “week”, “weekday”, “day”, “hour” , “minute”, “quarter”, “month” με χρήση του statetime ώστε να αντλήσουμε και να χρησιμοποιήσουμε τις πληροφορίες αυτές, αφού το statetime είναι τύπου datetime και δεν μπορεί να χρησιμοποιηθεί σε κατηγοριοποίηση ή συσταδοποίηση. Για τις υπόλοιπες κενές τιμές στις στήλες ”humidity_in”, ”humidity_out”, ”temp_int”, ”temp_out”, συμπληρώνουμε τα κενά υπολογίζοντας αρχικά τον μέσο όρο κάθε μέρας με χρήση της στήλης “day” και αντικαθιστώντας τον στις κενές τιμές της συγκεκριμένης ημέρας. Παρατηρήσαμε όμως ότι έλειπαν τιμές για κάποιες μέρες σε κάποιες περιπτώσεις, έτσι αυτές τις τιμές τις συμπληρώνουμε υπολογίζοντας τον μέσο όρο ολόκληρης της στήλης. Επιπλέον χρησιμοποιώντας την στήλη “hour”

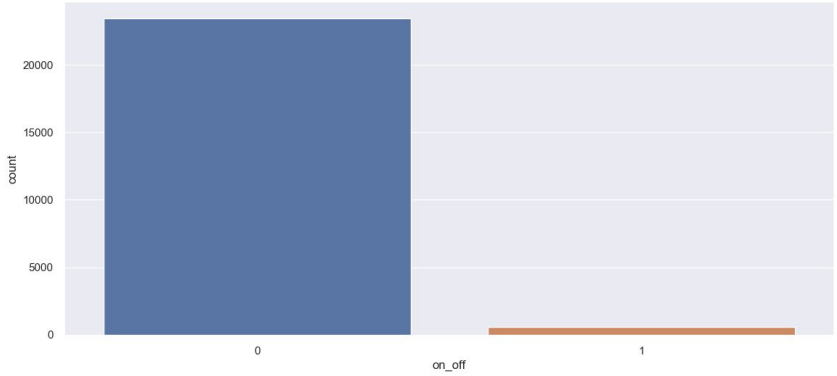
φτιάξαμε κάδους οι οποίοι αντιστοιχούσαν σε ['Morning','Afternoon','Evening', 'Night'] σπάζοντας την μέρα σε 4 κομμάτια [0, 5, 13, 17, 24]. Με την ίδια λογική αλλά χρησιμοποιώντας την στήλη “temp_out” φτιάξαμε κάδους οι οποίοι αντιστοιχούσαν σε ['Freezing','Cold','Warm','Hot','Very Hot'] σπάζοντας τις τιμές των θερμοκρασιών στα 5, [-10, 0, 15, 25, 33, 50] . Στην συνέχεια φτιάξαμε άλλη μια νέα στήλη, την “in_office” στην οποία βάζαμε 0 εάν δεν ήταν κάποιος στο γραφείο ενώ αντίστοιχα 1 εάν κάποιος ήταν στο γραφείο. Θεωρούμε ότι θα ήταν κάποιος στο γραφείο εάν το ac_pow_power + ac_pow_current + motion_detected ήταν μεγαλύτερο από 1 ενώ διαφορετικά στο γραφείο δεν ήταν κανείς.

Μετά την απαραίτητη προετοιμασία των δεδομένων, χρησιμοποιήσαμε την βιβλιοθήκη matplot ώστε να οπτικοποιήσουμε τα ολοκληρωμένα δεδομένα και να βγάλουμε κάποια αρχικά συμπεράσματα μέσω από γραφικές παραστάσεις.



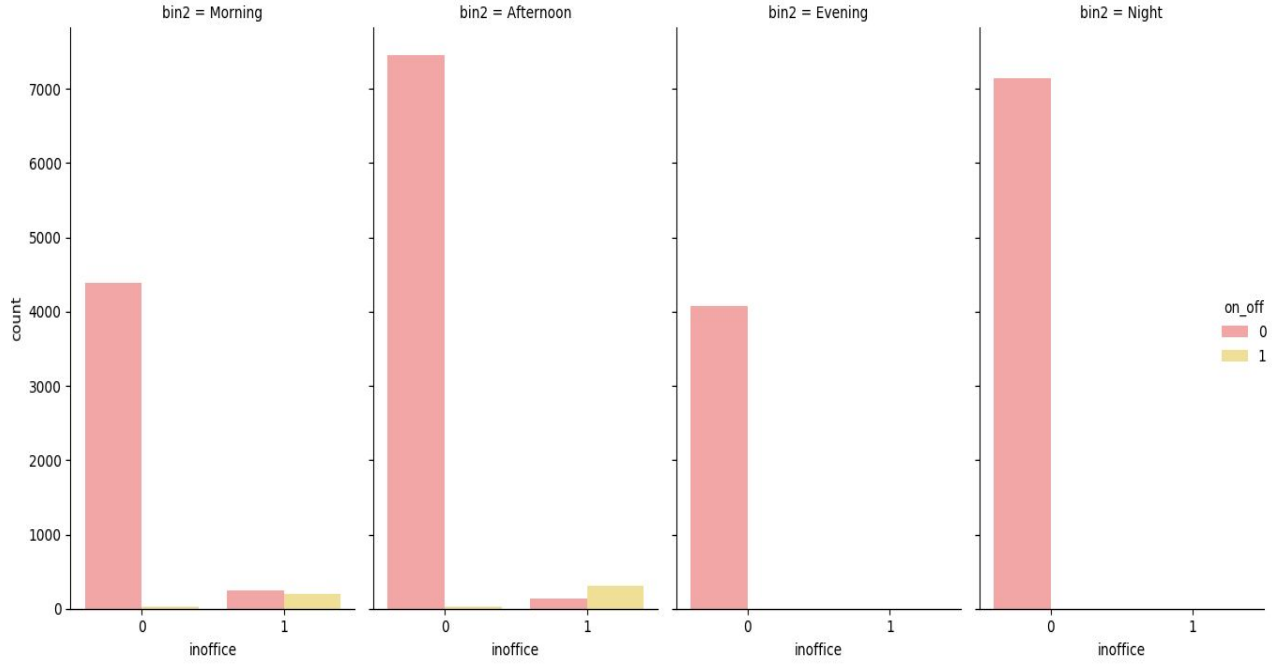
Σχήμα 1 Correlation Heatmap

Στο σχήμα 1 παρουσιάζεται ένα heatmap το οποίο δείχνει τις συσχετίσεις ανάμεσα στις τιμές των δεδομένων.



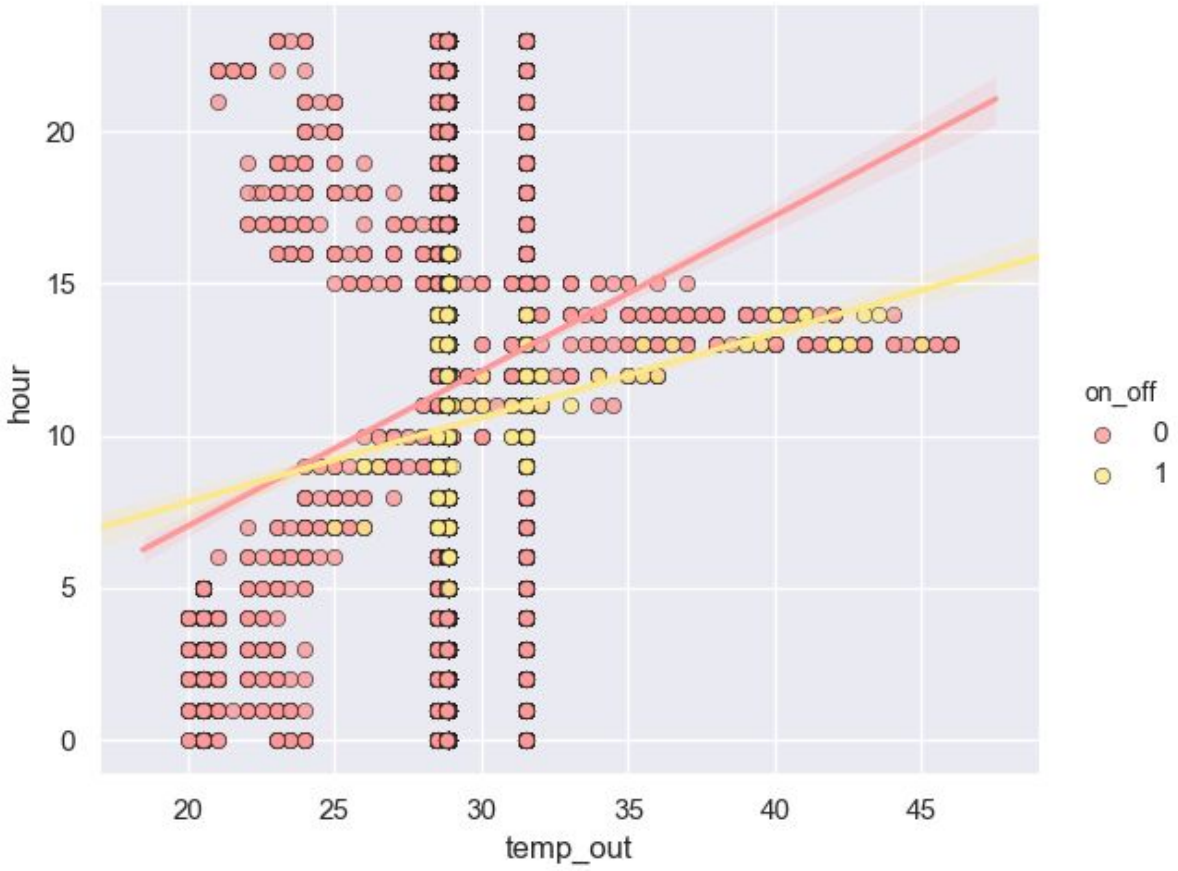
Σχήμα 2 Bar plot Count On_off

Στο σχήμα 2 παρουσιάζεται ένα Bar plot στο οποίο μπορούμε να δούμε πόσες τιμές on (1) υπάρχουν και πόσα off(0). Είναι φανερό πως στις περισσότερες μετρήσεις που καταγράφηκαν το κλιματιστικό ήταν κλειστό.



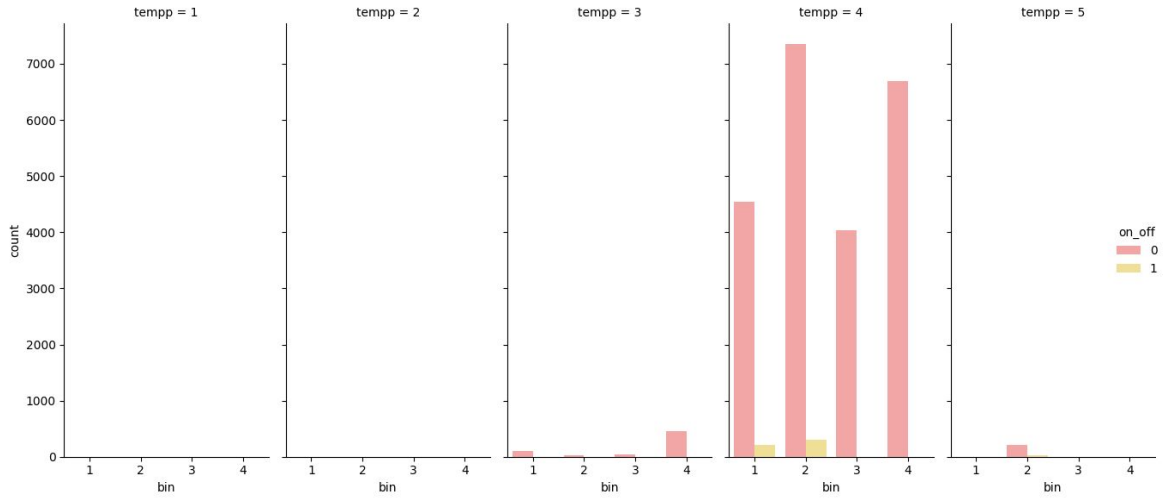
Σχήμα 3 Bar Plot

Στο σχήμα 3 παρουσιάζεται ένα bar plot στο οποίο απεικονίζεται με ροζ χρώμα πότε το κλιματιστικό ήταν κλειστό(off) και με μπεζ χρώμα πότε το κλιματιστικό ήταν ανοιχτό(on). Επίσης στον άξονα των X με τιμές 0 και 1 φαίνεται αν ήταν στο γραφείο(1) ή όχι(0) για τα 4 διαστήματα της ημέρας (Morning,Afternoon,Evening,Night). Παρατηρείται ότι κάποια πρωινά (Morning) και μεσημέρια (Afternoon) που ήταν στο γραφείο (τιμές 1 στον άξονα των X) άλλες φορές το κλιματιστικό ήταν κλειστό (ροζ μπάρα) και άλλες ανοιχτό(μπεζ μπάρα). Επίσης παρατηρείται ότι απόγευμα(Evening) και νύχτα(Night) το κλιματιστικό δεν άνοιξε καθόλου.



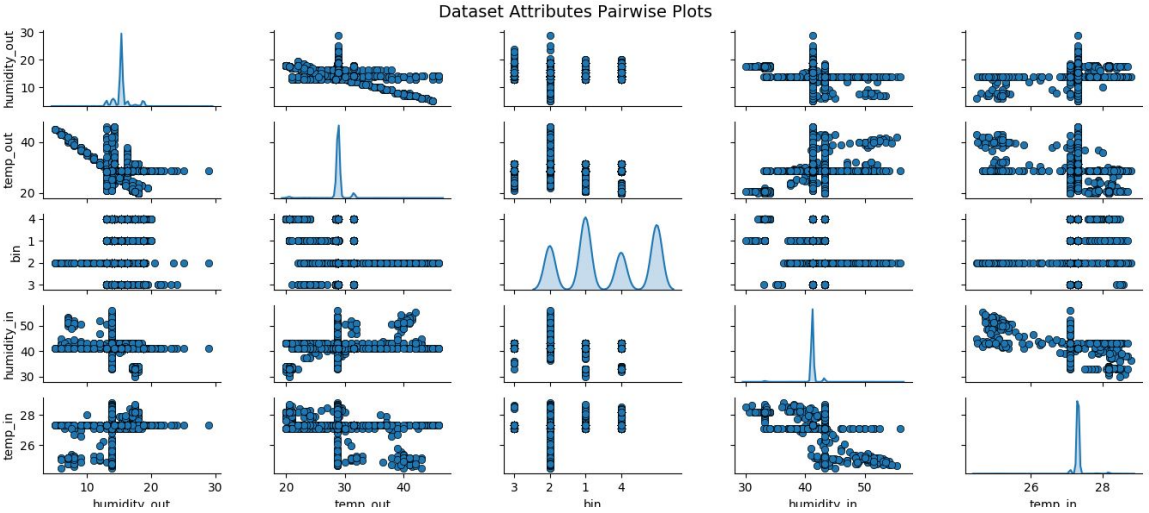
Σχήμα 4 Scatter Plot

Στο σχήμα 4 παρουσιάζεται η κατανομή των δειγμάτων σε σχέση με τον χρόνο και την εξωτερική θερμοκρασία, ενώ βάσει του χρώματος φαίνεται επίσης αν το κλιματιστικό είναι ανοιχτό ή κλειστό. Παρατηρείται ότι το κλιματιστικό είναι ανοιχτό κατά τις ώρες 6.00-16.00 με αντίστοιχη εξωτερική θερμοκρασία 25-42. Για το χρονικό αυτό διάστημα όπου παρατηρείται αυτή η μεταβολή της θερμοκρασίας, έχει εξαχθεί και το regression για τα αντίστοιχα σημεία.



Σχήμα 5 Bar Plot

Στο σχήμα 5 παρουσιάζεται ένα bar plot στο οποίο απεικονίζεται με ροζ χρώμα πότε το κλιματιστικό ήταν κλειστό(off) και με μπεζ χρώμα πότε το κλιματιστικό ήταν ανοιχτό(on). Επίσης στον άξονα των X με τιμές 1,2,3 και 4 φαίνονται τα 4 διαστήματα της ημέρας Morning,Afternoon,Evening και Night αντίστοιχα για τις 5 ζώνες εξωτερικής θερμοκρασίας (Freezing, Cold, Warm, Hot, Very Hot).Παρατηρείται ότι το κλιματιστικό ήταν ανοιχτό(μπεζ μπάρα) μόνο πρωινά(τιμή 1 στον άξονα τον X) και μεσημέρια(τιμή 2 στον άξονα τον X) όταν η εξωτερική θερμοκρασία ήταν “Hot” (από 25 έως 33 βαθμούς).



Σχήμα 6 Pairwise plots

Στο σχήμα 6 παρουσιάζονται συγκεντρωτικά οι κατανομές των φυσικών χαρακτηριστικών του dataset, όπως θερμοκρασία κλπ, σε σχέση με τα υπόλοιπα φυσικά χαρακτηριστικά όπως υγρασία, χρόνος κλπ.

Β) Κατηγοριοποίηση/Classification

Για να γίνει η κατηγοριοποίηση έπρεπε να οριστεί αρχικά το τι πρέπει να μαντέψει το μοντέλο, δηλαδή τον στόχο μας. Ο στόχος μας στην συγκεκριμένη εργασία είναι να μαντέψει το μοντέλο πότε το κλιματιστικό είναι on ή off. Είναι προφανές ότι έχουμε να κάνουμε με binary κατηγοριοποίηση μιας και οι τιμές που έχουμε να μαντέψουμε είναι μόνο 2 και μπορούν να χαρακτηριστούν με 0 ή 1. Αρχικά έπρεπε να γίνει η εκπαίδευση και η αξιολόγηση του μοντέλου και αργότερα οι προβλέψεις χρησιμοποιώντας τον καλύτερο δυνατό αλγόριθμο. Χρησιμοποιήθηκε το 20% ολόκληρου το dataset για testing και το υπόλοιπο 80% για training.

Για τα δεδομένα της εργασίας θέσαμε σαν στόχο-target την στήλη “on_off” και σαν features τις στήλες 'bin',’temp’,'motion_detected','humidity_in','humidity_out','t
emp_out','temp_in'. Όπου bin είναι η στήλη με τους αριθμούς των κάδων που αναφέρονται στα μέρη της ημέρας “Morning”,”Afternoon”,Evening”,Night ενώ όπου temp είναι η στήλη με τους αριθμούς των κάδων που αναφέρονται στους γενικότερους ορισμούς των θερμοκρασιών “Freezing”, ”Cold”, “Warm”, “Hot”, “Very Hot”.
Για την κατηγοριοποίηση χρησιμοποιήθηκαν κάποιοι απο τους αλγόριθμους κατηγοριοποίησης οι οποίοι και αξιολογήθηκαν με τα εργαλεία Confusion Matrix και Classification Report , η δε αξιολόγησή τους βάσει αυτών των εργαλείων έγινε και ξεχωριστά για τον καθένα αλλά και συγκριτικά για όλους τους αλγορίθμους που χρησιμοποιήθηκαν.

1) Confusion Matrix :

True Positive : actual = 1 predict = 1 (TP)
False Positive : actual = 0 predict = 1 (FP)
False Negative : actual = 1 predict = 0 (FN)
True Negative : actual = 0 predict = 0 (TN)

Όπως γνωρίζουμε, το Confusion Matrix, στην απλή περίπτωση του binary classification, μας δίνει έναν πίνακα που αναπαριστά τις αριθμητικές τιμές των TP,FP,FN,TN τα οποία εξηγούνται παραπάνω. Έτσι προκύπτουν οι εκτιμήσεις του αλγορίθμου και ποιες από αυτές ήταν False ή True, συνολικά για τα δείγματα που εξέτασε ο αλγόριθμος. Αυτό μας βοηθά σε μια πρώτη εκτίμηση για το πόσο καλά λειτούργησε ο αλγόριθμος που χρησιμοποιήσαμε, πχ υπολογίζοντας το ποσοστό των True ή False σε κάθε binary χαρακτηριστικό δηλ τα Positive,Negative.

2) Classification Reports

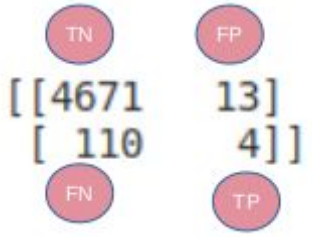
Όπως γνωρίζουμε, το Classification Report αποτελεί έναν πίνακα ο οποίος αναπαριστά μετρικές σχετικά με τις επιδόσεις του αλγορίθμου. Οι μετρικές αυτές υπολογίζονται από τα αριθμητικά στοιχεία που εντοπίζονται στο Confusion Matrix. Συγκεκριμένα οι μετρικές αυτές είναι: precision, recall, f1-score.

- precision,p: Αποτελεί για κάποιο binary χαρακτηριστικό, το ποσοστό των ορθών προβλέψεων που έκανε ο αλγόριθμος στο σύνολο των δειγμάτων που προβλέπει αρχικά ως τέτοιο binary χαρακτηριστικό.
- recall,r: Αποτελεί για κάποιο binary χαρακτηριστικό, το ποσοστό των ορθών προβλέψεων που έκανε ο αλγόριθμος στο σύνολο των πραγματικών δειγμάτων που έχουν το συγκεκριμένο binary χαρακτηριστικό.
- f1-score: ορίζεται ως ο αρμονικός μέσος των precision και recall, f1 = 1/p + 1/r

Αλγόριθμοι Κατηγοριοποίησης

1)Decision Tree

Confusion Matrix :



Με βάση την θεωρία για τους Confusion Matrix βλέπουμε ότι το μοντέλο μας μαντεύει σωστά 4671/4684 off ενώ 13/4684 φορές μαντεύει λάθος ενώ αντίστοιχα μαντεύει τις 110/114 φορές λάθος ότι το κλιματιστικό είναι ανοιχτό και 4/114 φορές πετυχαίνει την πρόβλεψη ότι το κλιματιστικό είναι ανοιχτό. Αυτό ίσως μπορεί να το αιτιολογεί το γεγονός ότι είχαμε πολύ λίγες μετρήσεις με το κλιματιστικό ανοιχτό και έτσι το μοντέλο μας το βρίσκει δυσκολότερο να κάνει σωστές προβλέψεις για ανοικτό κλιματιστικό.

Classification Report:

	precision	recall	f1-score	support
class 0	0.98	1.00	0.99	4684
class 1	0.24	0.04	0.06	114
accuracy			0.97	4798
macro avg	0.61	0.52	0.52	4798
weighted avg	0.96	0.97	0.97	4798

2)Gaussian Naive Bayes

Είναι βασισμένος στο θεώρημα του Bayes και υπολογίζει την πιθανότητα να έχει κάποια σύνδεση ένα feature με το target και επιλέγει αυτά με την μεγαλύτερη πιθανότητα.

Confusion Matrix:

```
Guassian CM
[[4473  211]
 [ 103   11]]
```

Με βάση την θεωρία για τους Confusion Matrix βλέπουμε ότι το μοντέλο μας μαντεύει σωστά 4473/4684 off ενώ 211/4684 φορές μαντεύει λάθος ενώ αντίστοιχα μαντεύει τις 103/114 φορές λάθος ότι το κλιματιστικό είναι ανοιχτό και 11/114 φορές πετυχαίνει την πρόβλεψη ότι το κλιματιστικό είναι ανοιχτό.

Εδώ παρατηρούμε ότι ο αλγόριθμος αυτός είχε λίγο καλύτερο ποσοστό επιτυχίας στο να μαντεύει ότι το κλιματιστικό είναι ανοιχτό ενώ χαμηλότερο όταν έπρεπε να μαντέψει ότι το κλιματιστικό ήταν κλειστό.

Classification Report:

	precision	recall	f1-score	support
class 0	0.98	0.96	0.97	4684
class 1	0.06	0.10	0.07	114
accuracy			0.94	4798
macro avg	0.52	0.53	0.52	4798
weighted avg	0.96	0.94	0.95	4798

3)K-Nearest Neighbors

Confusion Matrix :

```
KNN CM
[[4676    8]
 [ 110    4]]
```

Με βάση την θεωρία για τους Confusion Matrix βλέπουμε ότι το μοντέλο μας μαντεύει σωστά 4679/4684 off ενώ 8/4684 φορές μαντεύει λάθος ενώ αντίστοιχα μαντεύει τις 110/114 φορές λάθος ότι το κλιματιστικό είναι ανοιχτό και 4/114 φορές πετυχαίνει την πρόβλεψη ότι το κλιματιστικό είναι ανοιχτό.

Έτσι συμπεραίνουμε ότι έχει καλύτερο ποσοστό επιτυχίας σε σχέση με τον GaussianNB ενώ έχει περίπου τα ίδια ποσοστά επιτυχίας με τον Decision Tree.

Classification Report:

	precision	recall	f1-score	support
class 0	0.98	0.99	0.99	4684
class 1	0.23	0.06	0.10	114
accuracy			0.97	4798
macro avg	0.60	0.53	0.54	4798
weighted avg	0.96	0.97	0.97	4798

4)LinearSVC

Προσπαθεί να μοιράσει τα δεδομένα σε διαφορετικά sets ώστε να βρει το καλύτερο δυνατό ‘grouping’ των διαφορετικών κλάσεων.

Confusion Matrix:

```
SVC CM
[[4684    0]
 [ 114    0]]
```

Με βάση την θεωρία για τους Confusion Matrix βλέπουμε ότι το μοντέλο μας μαντεύει σωστά 4684/4684 off ενώ 0/4684 φορές μαντεύει λάθος ενώ αντίστοιχα μαντεύει τις 114/114 φορές λάθος ότι το κλιματιστικό είναι ανοιχτό και 0 φορές πετυχαίνει την πρόβλεψη ότι το κλιματιστικό είναι ανοιχτό.
Ο αλγόριθμος αυτός έχει πολύ καλά ποσοστά στο να μαντεύει ότι το κλιματιστικό είναι κλειστό ενώ έχει το χειρότερο ποσοστό επιτυχίας στο να μαντέψει ότι το κλιματιστικό είναι ανοιχτό.

Classification Report:

	precision	recall	f1-score	support
class 0	0.98	1.00	0.99	4684
class 1	0.00	0.00	0.00	114
accuracy			0.98	4798
macro avg	0.49	0.50	0.49	4798
weighted avg	0.95	0.98	0.96	4798

Συμπεράσματα Αλγορίθμων Κατηγοριοποίησης:

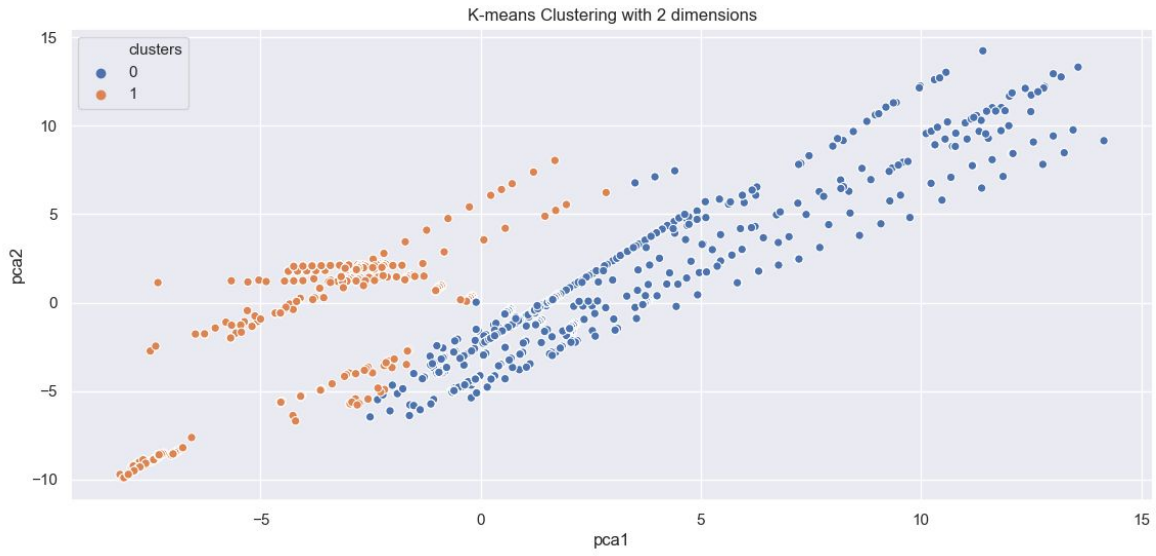
	Standard	Standard	Standard	Standard	Standard
1	Statistic Measure	Decision Tree	Naive Bayes	LinearSVC	KNN
2	Accuracy	0.9739	0.9346	0.9762	0.9754
3	Precision	0.1765	0.0495	0.0	0.9754
4	Recall	0.5117	0.5257	0.5	0.5167
5	F-Measure	0.9644	0.9447	0.9645	0.9656

Έχοντας μπροστά μας αυτά τα στατιστικά αποτελέσματα ενώ επίσης παρατηρώντας το αρχείο Report.csv με τις μετρικές ανα κάθε αλγόριθμο, θεωρήθηκε ότι καλύτερος για το μοντέλο προβλεψης θα είναι ο KNN και έτσι χρησιμοποιήθηκε για την πρόβλεψη των τελικών αποτελεσμάτων.

Γ) Άλλες τεχνικές ανάλυσης - Συσταδοποίηση/Clustering

Στην συσταδοποίηση καθώς δεν υπάρχουν αρχικές κλάσεις αλλά προσπαθούν να οριστούν με χρήση αλγορίθμων συσταδοποίησης, προσπαθήσαμε μέσω γραφικών να δούμε εάν μπορούμε να εξάγουμε κάποιο συμπέρασμα χρησιμοποιώντας τον αλγόριθμο KMeans και κατα πόσο ο αλγόριθμος είναι αποδοτικός με τα δικά μας δεδομένα.
Επειδή τα δεδομένα μας είναι πολυδιάστατα χρησιμοποιήσαμε PCA ώστε να τα φέρουμε σε μορφή δύο διαστάσεων ώστε να μπορέσουν να ζωγραφιστούν όλα στο plot.

Σχήμα 7 PCA graph before KMeans Clustering

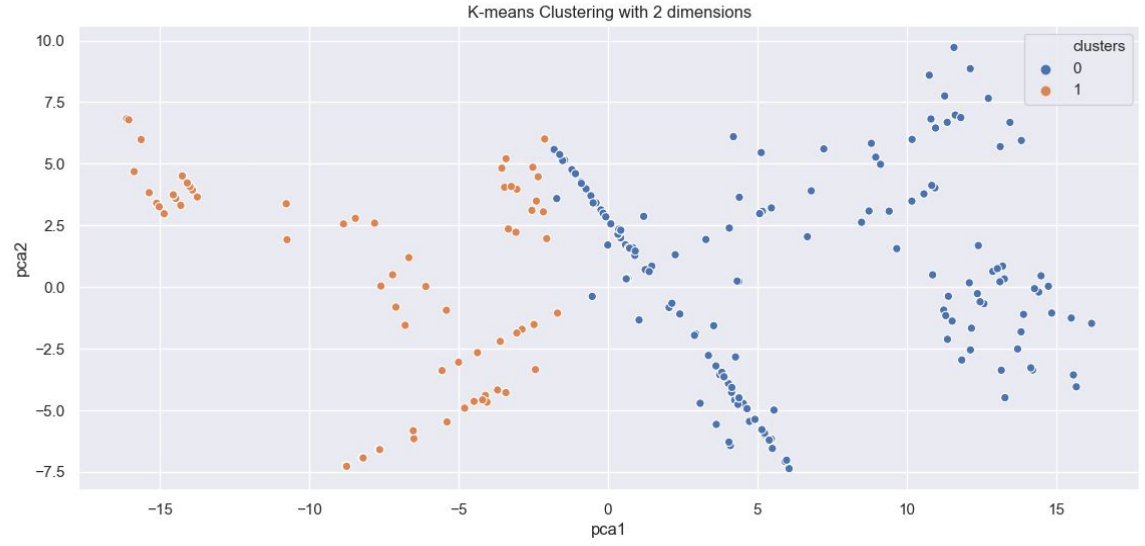


Σχήμα 8 PCA graph K-Means Clustering

Στο σχήμα 7 Παρουσιάζεται μια δισδιάστατη αναπαράσταση των δεδομένων πριν την συσταδοποίηση με βάση της στήλης on-off. Με χρώμα πορτοκαλί φαίνονται τα “on” ενώ με χρώμα μπλε τα “off”.
Στο σχήμα 8 Παρουσιάζεται μια δισδιάστατη αναπαράσταση των δεδομένων μετά την συσταδοποίηση με χρήση του αλγορίθμου K-Means, δημιουργώντας δύο clusters.
Παρατηρήσαμε ότι τα περισσότερα “on” έχουν μπει στο ίδιο cluster μετα την χρήση του αλγορίθμου συσταδοποίησης.

Πειραματικά στην περίπτωση που στα αρχικά δεδομένα χρησιμοποιούσαμε drop με tresh=3 ώστε να αποσυρθούν οι στήλες που έχουν περισσότερες από 3 nan τιμές οι αντίστοιχες γραφικές που προέκυψαν πρίν και μετά την συσταδοποίηση ήταν οι παρακάτω.

Σχήμα 9 Before Clustering Drop Tresh = 3



Σχήμα 10 K-Means Clustering Drop Tresh = 3

Συμπεράσματα από Αλγόριθμο Συσταδοποίησης :

Ο KMeans στα πρώτα σχήματα φαίνεται ότι δουλεύει σχετικά καλά αφού μπορεί έστω να βάλει σε διαφορετικά clusters τα on-off, αλλα αυτό δεν τον καθιστά τον καλύτερο και μάλλον δεν είναι ο κατάλληλος για τα συγκεκριμένα δεδομένα.

Δ) Δεδομένα Ελέγχου

Τα δεδομένα ελέγχου επεξεργάστηκαν με τον ίδιο τρόπο που επεξεργάστηκαν και τα δεδομένα εκπαίδευσης όπως αναφέρεται στο (Α).
Για την δημιουργία του .csv αρχείου των δεδομένων ελέγχου υλοποιήθηκε ξεχωριστό αρχείο κώδικα “test_file.py” το οποίο αφού τρέξει δημιουργεί το αρχείο “data_test.csv” το οποίο χρησιμοποιήθηκε σαν test set στην κατηγοριοποίηση για να προβλέψουμε τα τελικά μας αποτελέσματα.
Τα τελικά αποτελέσματα καταγράφονται στο αρχείο ac_statetime_test.csv για κάθε λεπτό όπως ζητήθηκε.

Ε) Γενικές Παρατηρήσεις

Μετά την προ-επεξεργασία των δεδομένων και μετά την χρήση των διαφόρων τεχνικών και αλγορίθμων στα συγκεκριμένα δεδομένα μπορούμε να πούμε ότι τα δεδομένα δεν έδιναν εύκολα τις πληροφορίες που χρειάζεται ένα μοντέλο για να κάνει τις καλύτερες δυνατές προβλέψεις. Πιό συγκεκριμένα τα δεδομένα εκπαίδευσης περιείχαν 4684 - off και μόνο 114 - on αυτό από μόνο του μπορούμε να πούμε ότι έκανε τις προβλέψεις για το 1 αρκετά πιο δύσκολες σε σχέση με τις προβλέψεις για το 0, αυτό μπορέσαμε να το δούμε και στα στατιστικά αποτελέσματα. Παράλληλα παρατηρήσαμε ότι το κλιματιστικό ήταν αναμμένο συνήθως σε θερμοκρασίες μεγαλύτερες απο 25. Ενώ λόγω του μήνα δεν μπορέσαμε να δούμε πολύ υψηλές θερμοκρασίες και πως λειτουργούσαν σε σχέση με το κλιματιστικό . Επίσης παρατηρήσαμε ότι το κλιματιστικό ήταν αναμμένο κυρίως πρωινά και μέχρι νωρίς το απόγευμα.