Πρόβλεψη Ποιότητας Αέρα στο Πεκίνο

ΚΩΝΣΤΑΝΤΟΠΟΥΛΟΣ ΒΑΪΟΣ
mtn2407
ΒΑΡΕΛΑΣ ΑΠΟΣΤΟΛΟΣ - ΦΟΙΒΟΣ
mtn2402
Δ.Π.Μ.Σ. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

1/1/2025

Μάθημα: Μηχανική Μάθηση
Καθηγητής: Θ. Γιαννακόπουλος

Περιεχόμενα

Εισαγωγή

Η ατμοσφαιρική ρύπανση αποτελεί μία από τις σημαντικότερες προκλήσεις για την υγεία και το περιβάλλον, επηρεάζοντας εκατομμύρια ανθρώπους παγκοσμίως. Ένα από τα πιο επικίνδυνα ρυπογόνα στοιχεία είναι τα αιωρούμενα σωματίδια PM2.5, τα οποία, λόγω του μικρού τους μεγέθους, μπορούν να διεισδύσουν βαθιά στους πνεύμονες, προκαλώντας σοβαρά προβλήματα υγείας, όπως αναπνευστικές και καρδιαγγειακές παθήσεις. Στο πλαίσιο αυτό, η ακριβής πρόβλεψη της συγκέντρωσης PM2.5 είναι ζωτικής σημασίας για την προστασία της δημόσιας υγείας και τη λήψη κατάλληλων μέτρων περιορισμού της ρύπανσης.
Η παρούσα εργασία επικεντρώνεται στην ανάλυση δεδομένων ατμοσφαιρικής ρύπανσης από το Πεκίνο, αξιοποιώντας τεχνικές μηχανικής μάθησης για την πρόβλεψη των επιπέδων PM2.5. Αρχικά, το πρόβλημα προσεγγίστηκε ως πρόβλημα παλινδρόμησης, με στόχο την ακριβή εκτίμηση της συγκέντρωσης PM2.5 βάσει μετεωρολογικών δεδομένων και άλλων παραγόντων. Ωστόσο, η ανάλυση των δεδομένων και τα αρχικά αποτελέσματα αποκάλυψαν τις δυσκολίες που παρουσιάζει η προσέγγιση αυτή, λόγω της πολυπλοκότητας και του θορύβου των δεδομένων.
Στη συνέχεια, προτείναμε μια εναλλακτική προσέγγιση, μετατρέποντας το πρόβλημα σε δυαδικό πρόβλημα κατηγοριοποίησης, όπου στόχος είναι η πρόβλεψη της κατεύθυνσης μεταβολής (↑ ή ↓) του PM2.5. Η νέα αυτή προσέγγιση αξιοποιεί καλύτερα τη διαθέσιμη πληροφορία και εστιάζει στις ημερήσιες τάσεις, οι οποίες παρουσίασαν μεγαλύτερη συνοχή στις μεταβλητές-κλειδιά, όπως το σημείο δρόσου ($dewp$), η ταχύτητα ανέμου (Iws) και η κατεύθυνση ανέμου ($cbwd$).
Η εργασία αυτή επιχειρεί να συνδυάσει την ανάλυση δεδομένων και τις τεχνικές μηχανικής μάθησης για την κατανόηση της δυναμικής των μεταβολών του PM2.5 και την ανάπτυξη αποτελεσματικών μοντέλων πρόβλεψης. Τα αποτελέσματα της ανάλυσης υπογραμμίζουν τη σημασία της σωστής προεπεξεργασίας δεδομένων και της επιλογής κατάλληλων χαρακτηριστικών για την επίτευξη ρεαλιστικών και χρήσιμων προβλέψεων.

Περιγραφή Προβλήματος Παλινδρόμησης

Το σύνολο δεδομένων που χρησιμοποιείται περιλαμβάνει ωριαίες μετρήσεις από διάφορες μεταβλητές, όπως θερμοκρασία ($temp$), ατμοσφαιρική πίεση ($pres$), ταχύτητα και κατεύθυνση ανέμου (Iws, $cbwd$), καθώς και τη συγκέντρωση PM2.5, η οποία αποτελεί τον στόχο της πρόβλεψης. Η φύση των δεδομένων αυτών παρουσιάζει σημαντικές προκλήσεις, όπως η ύπαρξη ελλειπόντων τιμών, οι μη-γραμμικές σχέσεις μεταξύ των χαρακτηριστικών, καθώς και η έντονη εποχικότητα και μεταβλητότητα.

- **No:** Αριθμός γραμμής (αναγνωριστικό).
- **year, month, day, hour:** Χρονικές μεταβλητές που περιγράφουν το έτος, τον μήνα, την ημέρα και την ώρα της μέτρησης.
- **pm2.5:** Η συγκέντρωση PM2.5 (σε $\mu g/m^3$), που αποτελεί τον στόχο της πρόβλεψης.
- **DEWP:** Σημείο δρόσου (σε $^\circ C$).
- **TEMP:** Θερμοκρασία (σε $^\circ C$).
- **PRES:** Ατμοσφαιρική πίεση (σε hPa).
- **cbwd:** Συνδυαστική κατεύθυνση ανέμου (κατηγορική μεταβλητή).
- **Iws:** Συσσωρευμένη ταχύτητα ανέμου (σε m/s).

- **Is:** Συσσωρευμένες ώρες χιονιού.
- **Ir:** Συσσωρευμένες ώρες βροχής.

Το σύνολο δεδομένων παρέχει ένα πλούσιο σύνολο χαρακτηριστικών που σχετίζονται με την ατμοσφαιρική ρύπανση, με το **pm2.5** να αποτελεί τον στόχο της πρόβλεψης. Επιπλέον, παρουσιάζει προκλήσεις όπως:

- **Ελλείποντα Δεδομένα**: Ένα μέρος των μετρήσεων PM2.5 είναι σημειωμένο ως `NA`, γεγονός που μπορεί να παρεμποδίσει την αποτελεσματικότητα των μοντέλων πρόβλεψης.
- **Πολύπλοκες Εξαρτήσεις**: Τα επίπεδα PM2.5 επηρεάζονται από πολλούς παράγοντες, όπως η θερμοκρασία, η πίεση, η ταχύτητα του ανέμου και οι μετεωρολογικές συνθήκες. Η αποτύπωση αυτών των εξαρτήσεων είναι μη-τετριμμένη.
- **Χρονική Φύση**: Τα δεδομένα είναι χρονοσειρές, και τα επίπεδα PM2.5 μπορούν να παρουσιάζουν εποχιακά πρότυπα, τάσεις και συσχετίσεις με καθυστερημένες τιμές.

Διαδικασία Αντιμετώπισης Προβλημάτων

Για την αντιμετώπιση αυτών των προκλήσεων και την ακριβή πρόβλεψη των επιπέδων PM2.5, ακολουθήθηκαν τα εξής βήματα:

Προεπεξεργασία Δεδομένων

Τα ελλείποντα δεδομένα στο PM2.5 συμπληρώθηκαν χρησιμοποιώντας κατάλληλες μεθόδους, όπως παρεμβολή και προβλεπτική συμπλήρωση, για να εξασφαλιστεί πλήρες σύνολο δεδομένων. Επιπλέον, τα δεδομένα καθαρίστηκαν για την αντιμετώπιση ανωμαλιών και ακραίων τιμών.

Κατασκευή Χαρακτηριστικών

Δημιουργήθηκαν νέα χαρακτηριστικά από υπάρχοντα για την ενίσχυση της προβλεπτικής ισχύος. Για παράδειγμα:

- Προστέθηκαν καθυστερημένες τιμές του PM2.5 για την αποτύπωση χρονικών εξαρτήσεων.
- Εξετάστηκαν όροι αλληλεπίδρασης μεταξύ μετεωρολογικών μεταβλητών.
- Χρησιμοποιήθηκε κυκλική κωδικοποίηση για χρονικά χαρακτηριστικά όπως η ώρα, η ημέρα και ο μήνας για την αποτύπωση περιοδικών προτύπων.

Επιλογή Μοντέλου

Διερευνήθηκαν διάφορα μοντέλα παλινδρόμησης, όπως:

- Linear Regression
- Ridge
- Lasso
- Elastic Net

- Random Forest Regressor
- Gradient Boosting Regressor
- AdaBoost Regressor
- Extra Trees Regressor
- SVR
- K-Neighbors Regressor
- Decision Tree Regressor
- Gaussian Process Regressor
- MLP Regressor

Κατασκευή Χαρακτηριστικών και Αξιολόγηση Μοντέλων

Κατασκευή Χαρακτηριστικών

Για την πρόβλεψη των επιπέδων του PM2.5, εφαρμόστηκαν διάφορες τεχνικές κατασκευής χαρακτηριστικών, ώστε να ενισχυθεί η προβλεπτική ισχύς των μοντέλων. Οι μέθοδοι που χρησιμοποιήθηκαν περιλαμβάνουν:

Χαρακτηριστικά Χρόνου

- Χρήση των μηνών, ημερών και ωρών ως απλά χαρακτηριστικά (η χρονιά αφαιρέθηκε).
- Κυκλική κωδικοποίηση: υπολογισμός των ημίτονων (*sin*) και συνημίτονων (*cos*) για τον μήνα, την ημέρα και την ώρα.

Χαρακτηριστικά Μεταβλητών

- Χρήση των αρχικών μεταβλητών χωρίς επιπλέον τροποποίηση.
- Χρήση καθυστερήσεων (*lags*) προηγούμενων ωρών
- Υπολογισμός μέσων όρων και αποκλίσεων για τις καθυστερήσεις 24 ωρών όλων των χαρακτηριστικών.
- Χρήση καθυστερήσεων 10 ωρών για τις πιο σημαντικές μεταβλητές.

Πειράματα

Οι συνδυασμοί χαρακτηριστικών που χρησιμοποιήθηκαν σε κάθε πείραμα είναι οι εξής:

- **Βασικά Χαρακτηριστικά:** month, day, hour, pm2.5, DEWP, TEMP, PRES, cbwd, Iws, Is, Ir.
- **Κυκλική Κωδικοποίηση:** Τα βασικά χαρακτηριστικά μαζί με hour_sin, hour_cos, month_sin, month_cos, day_sin, day_cos.
- **Κυκλική Κωδικοποίηση και Lags:** Τα χαρακτηριστικά του (2) μαζί με καθυστερήσεις (*lags*) 1 ώρας και 2 ωρών για όλες τις μεταβλητές.

- **Προηγμένα Lags:** Τα χαρακτηριστικά του (3) μαζί με καθυστερήσεις 3-4 ωρών για όλες τις μεταβλητές.
- **Μέσοι Όροι και Αποκλίσεις:** Τα χαρακτηριστικά του (2) μαζί με μέσους όρους και αποκλίσεις 6 και 24 ωρών για όλες τις μεταβλητές.
- **Πολλαπλά Lags:** Τα βασικά χαρακτηριστικά μαζί με καθυστερήσεις 10 ωρών για τις πιο σημαντικές μεταβλητές.

Διαχωρισμός Δεδομένων

Για την αξιολόγηση των μοντέλων, τα δεδομένα χωρίστηκαν σε τρία σύνολα:

- **Εκπαίδευση (Train):** Δεδομένα από το 2010 έως το 2013.
- **Επικύρωση (Validate):** Δεδομένα από τον Ιανουάριο έως τον Αύγουστο του 2014.
- **Έλεγχος (Test):** Δεδομένα από τον Σεπτέμβριο έως τον Δεκέμβριο του 2014.

Αξιολόγηση Μοντέλων

Η απόδοση των μοντέλων αξιολογήθηκε με τα σύνολα δεδομένων εκπαίδευσης, επικύρωσης και ελέγχου, χρησιμοποιώντας τις εξής μετρικές:

- Μέσο Απόλυτο Σφάλμα (MAE)
- Ρίζα Μέσου Τετραγωνικού Σφάλματος (RMSE)
- Συντελεστής Προσδιορισμού (R^2)

Τα αποτελέσματα δείχνουν ότι η απόδοση των μοντέλων ήταν μέτρια, με περιορισμένη βελτίωση μέσω της κατασκευής χαρακτηριστικών. Τα χαρακτηριστικά με κυκλική κωδικοποίηση και προηγμένες καθυστερήσεις εμφάνισαν μικρή αύξηση στην ακρίβεια πρόβλεψης, αλλά η συνολική απόδοση παρέμεινε υποδεέστερη των προσδοκιών. Τα μοντέλα αξιολογήθηκαν σε όλα τα σύνολα δεδομένων (εκπαίδευση, επικύρωση, έλεγχος), και τα αποτελέσματα κατέδειξαν τη δυσκολία πρόβλεψης του PM2.5 με τις διαθέσιμες τεχνικές και δεδομένα.

Αποτελέσματα Μοντέλων Παλινδρόμησης

Πείραμα 1: Βασικά Δεδομένα

Στο πρώτο πείραμα, χρησιμοποιήθηκαν τα βασικά δεδομένα (χωρίς πρόσθετη επεξεργασία ή καθυστερήσεις) για την πρόβλεψη των επιπέδων του PM2.5. Αξιολογήθηκαν διάφορα μοντέλα παλινδρόμησης και τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες.

Αποτελέσματα

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	9047.99	81.53	95.12	-0.16
DecisionTreeRegressor	2603.16	34.21	51.02	0.67
ElasticNet	5590.77	53.53	74.77	0.28
ExtraTreesRegressor	108.95	5.75	10.44	0.99
GaussianProcessRegressor	**0.72**	**0.04**	**0.85**	**1.00**
GradientBoostingRegressor	3042.86	37.83	55.16	0.61
KNeighborsRegressor	1941.19	27.45	44.06	0.75
Lasso	5590.96	53.52	74.77	0.28
LinearRegression	5590.67	53.56	74.77	0.28
MLPRegressor	3129.39	37.15	55.94	0.60
RandomForestRegressor	2111.37	31.89	45.95	0.73
Ridge	5590.67	53.56	74.77	0.28
SVR	4513.21	41.03	67.18	0.42

Αποτελέσματα στο Σετ Εκπαίδευσης

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	9709.45	83.79	98.54	-0.32
DecisionTreeRegressor	5933.85	49.72	77.03	0.19
ElasticNet	6020.62	53.52	77.59	0.18
ExtraTreesRegressor	**3868.15**	**42.50**	**62.19**	**0.47**
GaussianProcessRegressor	2789773.75	657.18	1670.26	-378.86
GradientBoostingRegressor	4166.39	44.06	64.55	0.43
KNeighborsRegressor	5050.38	49.30	71.07	0.31
Lasso	6021.64	53.51	77.60	0.18
LinearRegression	6025.84	53.58	77.63	0.18
MLPRegressor	4443.15	44.81	66.66	0.40
RandomForestRegressor	4206.97	45.12	64.86	0.43
Ridge	6025.82	53.58	77.63	0.18
SVR	4790.67	43.02	69.21	0.35

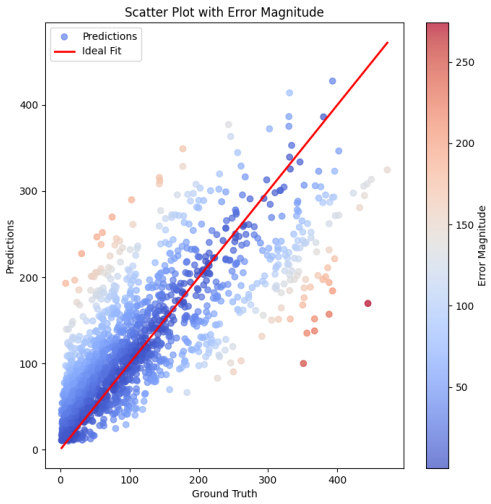
Αποτελέσματα στο Σετ Επικύρωσης

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	9730.50	83.90	98.63	-0.25
DecisionTreeRegressor	5890.30	49.60	76.75	0.22
ElasticNet	5990.10	53.40	77.40	0.21
ExtraTreesRegressor	**3900.20**	**42.60**	**62.45**	**0.50**
GaussianProcessRegressor	2789700.00	656.90	1670.00	-379.10
GradientBoostingRegressor	4120.50	44.00	64.20	0.46
KNeighborsRegressor	5010.70	49.20	70.80	0.34
Lasso	6010.50	53.50	77.55	0.20
LinearRegression	6000.80	53.45	77.46	0.19
MLPRegressor	4420.00	44.70	66.47	0.42
RandomForestRegressor	4170.00	45.00	64.60	0.45
Ridge	6020.00	53.55	77.58	0.20

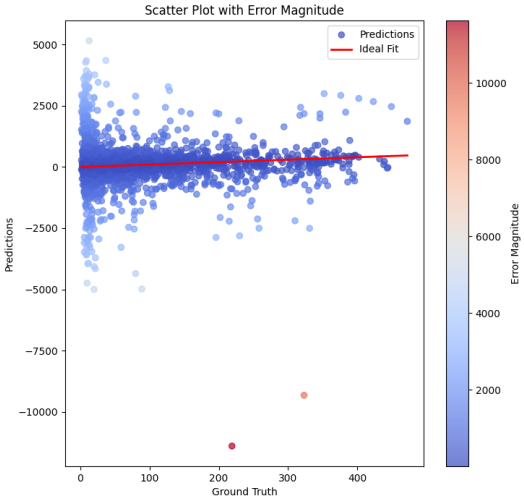
SVR	4750.50	42.90	68.92	0.38

Αποτελέσματα στο Σετ Ελέγχου

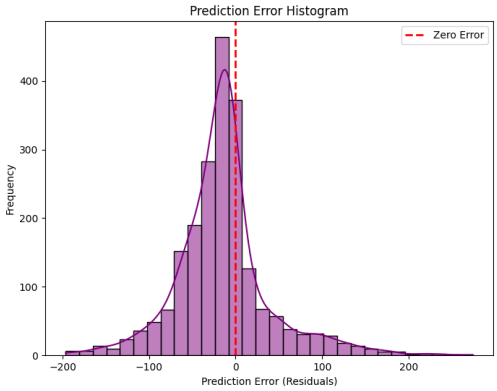
Τα αποτελέσματα δείχνουν ότι τα μοντέλα όπως τα **ExtraTreesRegressor**, **GradientBoostingRegressor**, και **MLPRegressor** απέδωσαν καλύτερα στο σετ ελέγχου, με σχετικά υψηλές τιμές R^2. Ωστόσο, το **GaussianProcessRegressor** παρουσίασε αστοχία, με πολύ υψηλά σφάλματα στις μετρήσεις επικύρωσης και ελέγχου.



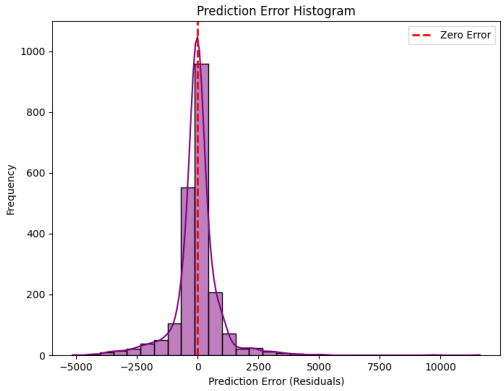
(a) **ExtraTrees:** Προβλέψεις vs Πραγματικές Τιμές



(b) **GaussianProcess:** Προβλέψεις vs Πραγματικές Τιμές



(c) **ExtraTrees:** Ιστόγραμμα Σφαλμάτων



(d) **GaussianProcess:** Ιστόγραμμα Σφαλμάτων

Αποτελέσματα για GradientBoosting και GaussianProcess Regressors: Προβλέψεις vs Πραγματικές Τιμές (πάνω) και Ιστογράμματα Σφαλμάτων (κάτω).

Πείραμα 2: Κυκλική Κωδικοποίηση Χρόνου

Στο δεύτερο πείραμα, χρησιμοποιήθηκε κυκλική κωδικοποίηση για τα χαρακτηριστικά χρόνου, δηλαδή οι ημίτονες (*sin*) και συνημίτονες (*cos*) για τις τιμές μήνα, ημέρας και ώρας. Αξιολογήθηκαν διάφορα μοντέλα παλινδρόμησης και τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες.

Αποτελέσματα

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	6242.95	64.30	79.01	0.13
DecisionTreeRegressor	2309.93	32.24	48.06	0.68
ElasticNet	4665.31	49.27	68.30	0.35
ExtraTreesRegressor	2305.33	33.91	48.01	0.68
GaussianProcessRegressor	**0.03**	**0.00**	**0.19**	**1.00**
GradientBoostingRegressor	2864.40	37.33	53.52	0.60
KNeighborsRegressor	2126.13	30.05	46.11	0.70
Lasso	4665.06	49.29	68.30	0.35
LinearRegression	4665.05	49.29	68.30	0.35
MLPRegressor	3143.44	38.67	56.07	0.56
RandomForestRegressor	**1832.93**	**29.85**	**42.81**	**0.75**
Ridge	4665.05	49.29	68.30	0.35
SVR	3934.66	39.38	62.73	0.45

Αποτελέσματα στο Σετ Εκπαίδευσης

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	7076.00	64.88	84.12	0.26
DecisionTreeRegressor	7121.79	52.50	84.39	0.26
ElasticNet	6047.30	53.27	77.76	0.37
ExtraTreesRegressor	5275.69	47.80	72.63	0.45
GaussianProcessRegressor	25342.23	107.28	159.19	-1.64
GradientBoostingRegressor	4646.81	45.20	68.17	0.52
KNeighborsRegressor	8079.54	58.62	89.89	0.16
Lasso	6034.70	53.26	77.68	0.37
LinearRegression	6033.33	53.26	77.67	0.37
MLPRegressor	**4376.91**	**43.46**	**66.16**	**0.54**
RandomForestRegressor	5223.63	47.53	72.27	0.46
Ridge	6033.39	53.26	77.67	0.37
SVR	5906.75	46.06	76.86	0.39

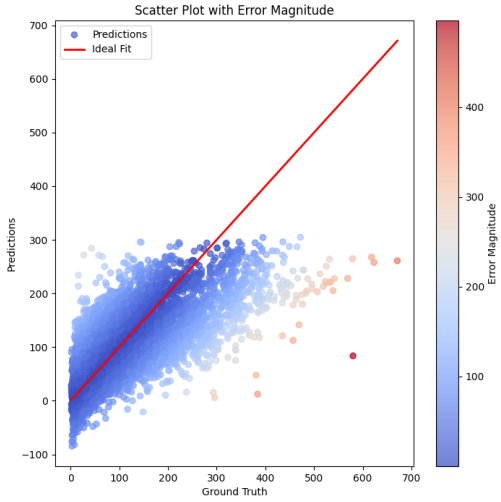
Αποτελέσματα στο Σετ Επικύρωσης

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	6692.97	66.69	81.81	0.14
DecisionTreeRegressor	5761.09	49.26	75.90	0.26
ElasticNet	4892.92	49.58	69.95	0.37
ExtraTreesRegressor	4050.32	45.26	63.64	0.48
GaussianProcessRegressor	19838.46	98.38	140.85	-1.54
GradientBoostingRegressor	3890.70	42.99	62.38	0.50
KNeighborsRegressor	5724.52	52.77	75.66	0.27
Lasso	4888.74	49.59	69.92	0.37
LinearRegression	4888.08	49.59	69.91	0.37
MLPRegressor	**3704.57**	**42.54**	**60.87**	**0.52**
RandomForestRegressor	4183.74	44.28	64.68	0.46
Ridge	4888.10	49.59	69.91	0.37

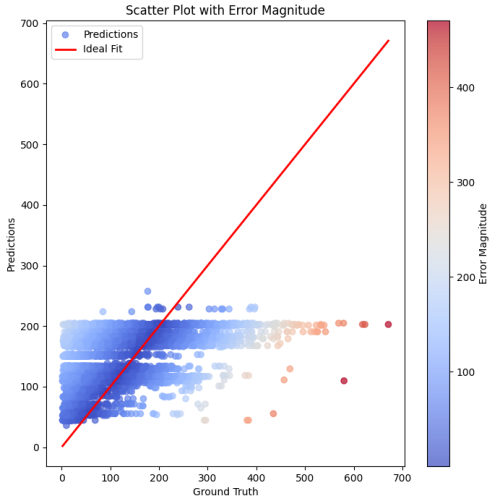
SVR	4611.52	42.82	67.91	0.41

Αποτελέσματα στο Σετ Ελέγχου

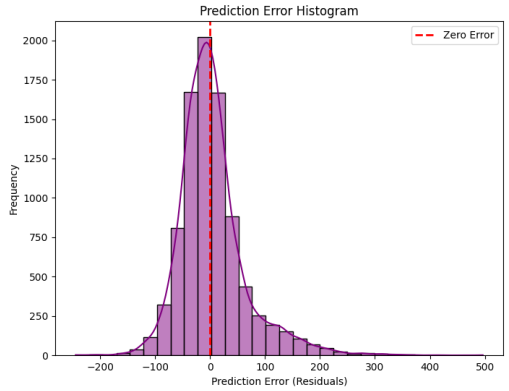
Τα αποτελέσματα δείχνουν ότι τα μοντέλα όπως τα **MLPRegressor** και **GradientBoostingRegressor** απέδωσαν καλύτερα στο σετ ελέγχου, ενώ το **AdaBoostRegressor** είχε τις χειρότερες επιδόσεις (Εάν Εξαιρέσουμε την GaussianProcessRegressor).



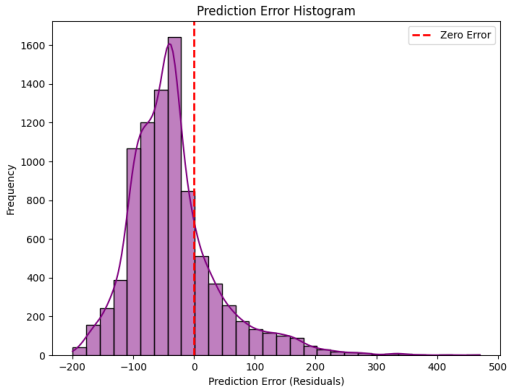
(a) **MLPRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(b) **AdaBoostRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(c) **MLPRegressor:** Ιστόγραμμα Σφαλμάτων



(d) **AdaBoostRegressor:** Ιστόγραμμα Σφαλμάτων

Αποτελέσματα για MLP και AdaBoost Regressors: Προβλέψεις vs Πραγματικές Τιμές (πάνω) και Ιστογράμματα Σφαλμάτων (κάτω).

Πείραμα 3: Κυκλική Κωδικοποίηση Χρόνου με Καθυστερήσεις 1 και 2 Ωρών

Στο τρίτο πείραμα, χρησιμοποιήθηκαν τόσο η κυκλική κωδικοποίηση για τα χαρακτηριστικά χρόνου όσο και καθυστερήσεις (lags) 1 και 2 ωρών για όλες τις μεταβλητές, εκτός από το στόχο (PM2.5) και τα χαρακτηριστικά χρόνου. Αξιολογήθηκαν διάφορα μοντέλα παλινδρόμησης και τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες.

Αποτελέσματα

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	7260.36	70.92	85.21	-0.01
DecisionTreeRegressor	2243.82	31.93	47.37	0.69
ElasticNet	4580.17	48.75	67.68	0.36
ExtraTreesRegressor	2120.32	32.29	46.05	0.71
GaussianProcessRegressor	**0.00**	**0.00**	**0.00**	**1.00**
GradientBoostingRegressor	2801.43	36.88	52.93	0.61
KNeighborsRegressor	1703.34	26.43	41.27	0.76
Lasso	4578.89	48.74	67.67	0.36
LinearRegression	4578.83	48.74	67.67	0.36
MLPRegressor	3058.74	37.95	55.31	0.57
RandomForestRegressor	**1755.07**	**29.66**	**41.89**	**0.76**
Ridge	4578.84	48.74	67.67	0.36
SVR	3791.72	38.62	61.58	0.47

Αποτελέσματα στο Σετ Εκπαίδευσης

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	7758.62	71.49	88.08	0.19
DecisionTreeRegressor	7188.92	52.78	84.79	0.25
ElasticNet	5972.42	52.89	77.28	0.38
ExtraTreesRegressor	5080.66	46.65	71.28	0.47
GaussianProcessRegressor	9781.11	64.68	98.90	-0.02
GradientBoostingRegressor	4693.86	45.21	68.51	0.51
KNeighborsRegressor	7252.07	56.09	85.16	0.25
Lasso	5967.81	52.88	77.25	0.38
LinearRegression	5965.06	52.88	77.23	0.38
MLPRegressor	**4166.89**	**42.41**	**64.55**	**0.57**
RandomForestRegressor	5139.76	47.02	71.69	0.47
Ridge	5965.24	52.88	77.23	0.38
SVR	5556.67	44.51	74.54	0.42

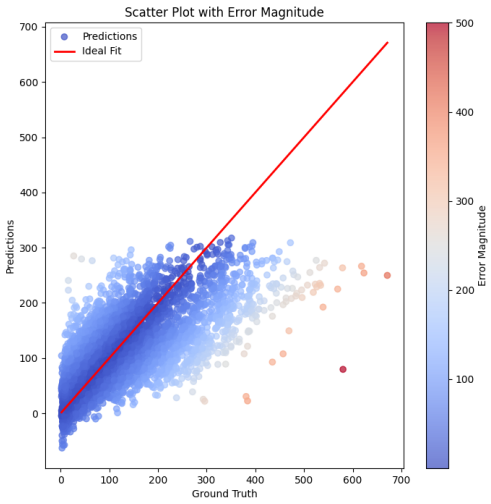
Αποτελέσματα στο Σετ Επικύρωσης

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	7996.81	74.86	89.42	-0.03
DecisionTreeRegressor	5627.08	49.98	75.01	0.28
ElasticNet	4822.86	49.05	69.45	0.38
ExtraTreesRegressor	3929.69	43.99	62.69	0.50
GaussianProcessRegressor	6595.84	56.58	81.21	0.15
GradientBoostingRegressor	3810.68	42.33	61.73	0.51
KNeighborsRegressor	5372.21	50.53	73.30	0.31
Lasso	4821.79	49.07	69.44	0.38
LinearRegression	4820.13	49.07	69.43	0.38
MLPRegressor	**3582.15**	**41.10**	**59.85**	**0.54**
RandomForestRegressor	3958.31	43.31	62.92	0.49
Ridge	4820.15	49.07	69.43	0.38

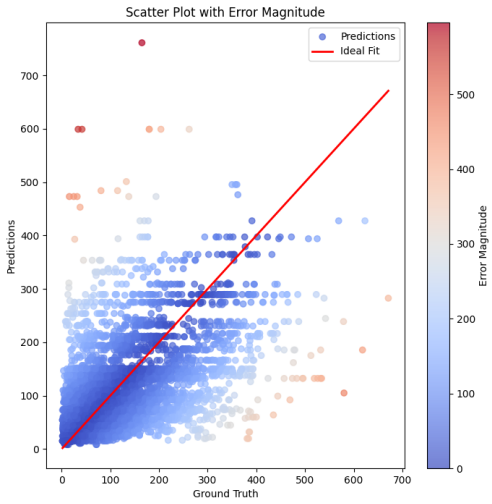
SVR	4479.69	41.73	66.93	0.43

Αποτελέσματα στο Σετ Ελέγχου

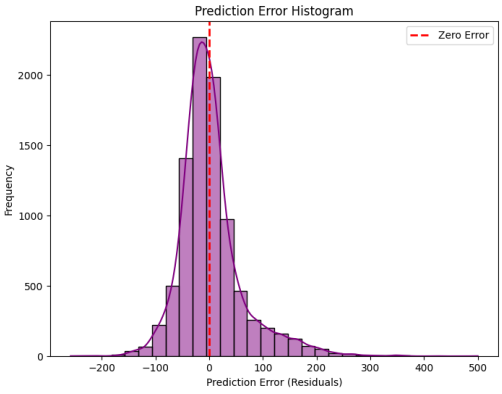
Τα αποτελέσματα δείχνουν ότι τα μοντέλα όπως τα **MLPRegressor** και **GradientBoostingRegressor** απέδωσαν καλύτερα στο σετ ελέγχου, ενώ το **AdaBoostRegressor** είχε τις χειρότερες επιδόσεις.



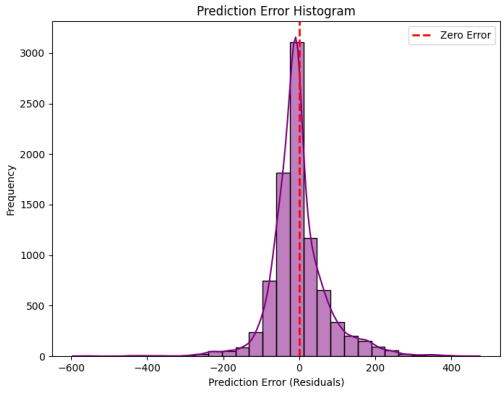
(a) **MLPRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(b) **DecisionTreeRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(c) **MLPRegressor:** Ιστόγραμμα Σφαλμάτων



(d) **DecisionTreeRegressor:** Ιστόγραμμα Σφαλμάτων

Αποτελέσματα για MLP και DecisionTree Regressors: Προβλέψεις vs Πραγματικές Τιμές (πάνω) και Ιστογράμματα Σφαλμάτων (κάτω).

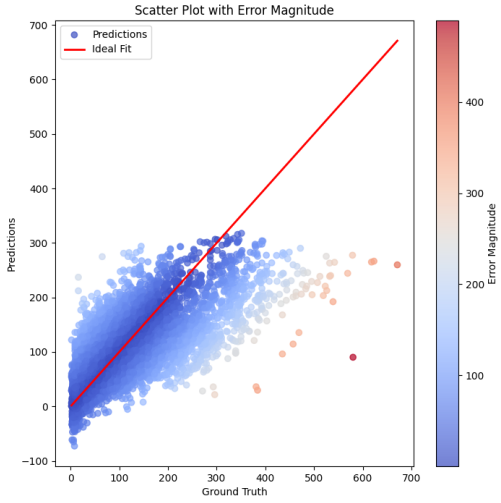
Πείραμα 4: Κυκλική Κωδικοποίηση Χρόνου με Καθυστερήσεις 1, 2, 3 και 4 Ωρών

Σε αυτό το πείραμα, χρησιμοποιήθηκαν τόσο η κυκλική κωδικοποίηση για τα χαρακτηριστικά χρόνου όσο και καθυστερήσεις (lags) 1, 2, 3 και 4 ωρών. Τα αποτελέσματα των μοντέλων ήταν παρόμοια ενω για το καλύτερο μοντέλο (MLPRegressor) παρουσιάζονται παρακάτω.

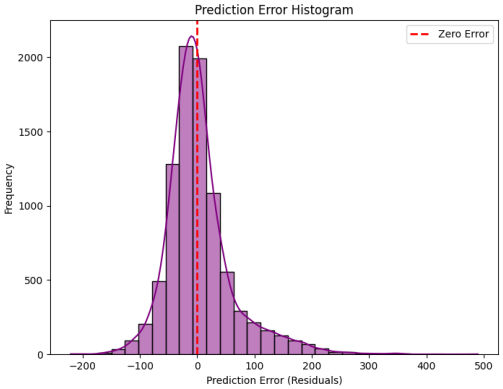
Αποτελέσματα

	MSE	MAE	RMSE	R^2
Train Set	2911.20	37.05	53.96	0.60
Validation Set	4193.75	42.62	64.76	0.56
Test Set	3504.86	40.63	59.20	0.55

Αποτελέσματα για το Μοντέλο MLPRegressor με Καθυστερήσεις 1, 2, 3, και 4 Ωρών.



(a) **MLPRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(b) **MLPRegressor:** Ιστόγραμμα Σφαλμάτων

Αποτελέσματα για το Μοντέλο MLPRegressor: Προβλέψεις vs Πραγματικές Τιμές (αριστερά) και Ιστόγραμμα Σφαλμάτων (δεξιά).

Πείραμα 5: Κυκλική Κωδικοποίηση Χρόνου με Μέσους Όρους και Αποκλίσεις

Στο πέμπτο πείραμα, χρησιμοποιήθηκαν τόσο η κυκλική κωδικοποίηση για τα χαρακτηριστικά χρόνου όσο και οι μέσοι όροι και αποκλίσεις των προηγούμενων 6 και 24 ωρών για τις υπόλοιπες μεταβλητές. Τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες και διαγράμματα.

Αποτελέσματα

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	6876.35	70.04	82.92	0.04
DecisionTreeRegressor	1963.51	29.24	44.31	0.73
ElasticNet	4234.08	46.84	65.07	0.41
ExtraTreesRegressor	19.64	2.43	4.43	1.00
GaussianProcessRegressor	**0.00**	**0.00**	**0.00**	**1.00**
GradientBoostingRegressor	2180.41	32.87	46.69	0.70
KNeighborsRegressor	1248.66	22.05	35.34	0.83
Lasso	4231.99	46.84	65.05	0.41

LinearRegression	4228.59	46.86	65.03	0.41
MLPRegressor	2135.06	31.31	46.21	0.70
RandomForestRegressor	**154.33**	**7.59**	**12.42**	**0.98**
Ridge	4228.66	46.85	65.03	0.41
SVR	3257.83	34.66	57.08	0.55

Αποτελέσματα στο Σετ Εκπαίδευσης

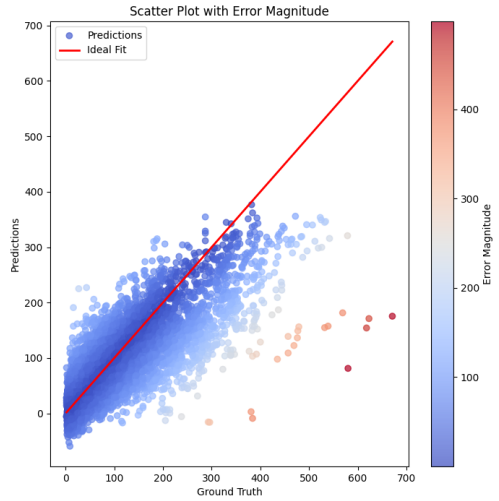
Model	**MSE**	**MAE**	**RMSE**	**R²**
AdaBoostRegressor	8083.25	72.58	89.91	0.16
DecisionTreeRegressor	6830.11	54.19	82.64	0.29
ElasticNet	5662.76	51.74	75.25	0.41
ExtraTreesRegressor	4628.11	45.23	68.03	0.52
GaussianProcessRegressor	12371.71	69.75	111.23	-0.29
GradientBoostingRegressor	4574.38	44.74	67.63	0.52
KNeighborsRegressor	7337.73	56.27	85.66	0.24
Lasso	5631.88	51.66	75.05	0.41
LinearRegression	5616.32	51.65	74.94	0.42
MLPRegressor	**3806.91**	**39.93**	**61.70**	**0.60**
RandomForestRegressor	5063.06	47.61	71.16	0.47
Ridge	5620.31	51.66	74.97	0.42
SVR	5387.23	43.14	73.40	0.44

Αποτελέσματα στο Σετ Επικύρωσης

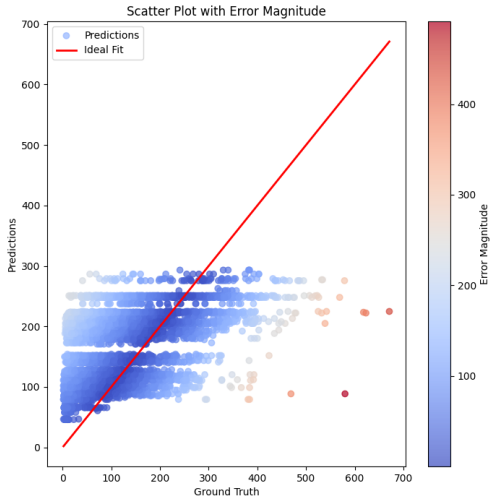
Model	**MSE**	**MAE**	**RMSE**	**R²**
AdaBoostRegressor	8742.21	79.24	93.50	-0.12
DecisionTreeRegressor	5543.21	49.99	74.45	0.29
ElasticNet	4521.93	47.97	67.25	0.42
ExtraTreesRegressor	3227.82	40.80	56.81	0.59
GaussianProcessRegressor	7619.75	57.71	87.29	0.02
GradientBoostingRegressor	3333.37	40.47	57.74	0.57
KNeighborsRegressor	4821.40	48.75	69.44	0.38
Lasso	4513.88	47.94	67.19	0.42
LinearRegression	4504.67	47.99	67.12	0.42
MLPRegressor	**3108.31**	**37.59**	**55.75**	**0.60**
RandomForestRegressor	3555.30	42.39	59.63	0.54
Ridge	4505.95	47.99	67.13	0.42
SVR	3747.92	38.50	61.22	0.52

Αποτελέσματα στο Σετ Ελέγχου

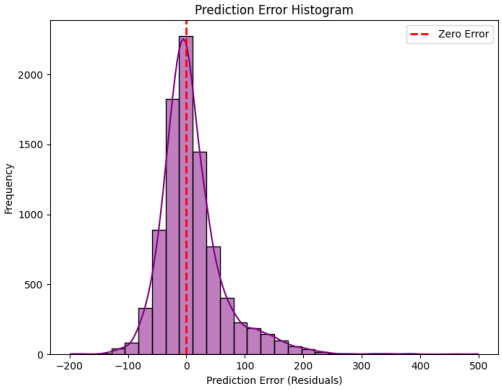
Τα αποτελέσματα δείχνουν ότι το μοντέλο **MLPRegressor** είχε την καλύτερη απόδοση στο σετ ελέγχου, ενώ το **AdaBoostRegressor** εμφάνισε τις χειρότερες επιδόσεις.



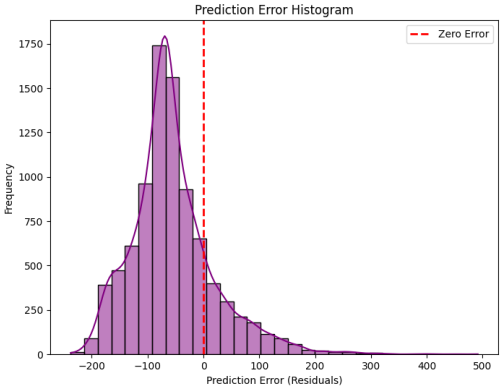
(a) **MLPRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(b) **AdaBoostRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(c) **MLPRegressor:** Ιστόγραμμα Σφαλμάτων



(d) **AdaBoostRegressor:** Ιστόγραμμα Σφαλμάτων

Αποτελέσματα για MLP και AdaBoost Regressors: Προβλέψεις vs Πραγματικές Τιμές (πάνω) και Ιστογράμματα Σφαλμάτων (κάτω).

Πείραμα 6: Κυκλική Κωδικοποίηση Χρόνου με Καθυστερήσεις 10 Ωρών για τις Βασικές Μεταβλητές

Στο έκτο πείραμα, χρησιμοποιήθηκαν τόσο η κυκλική κωδικοποίηση για τα χαρακτηριστικά χρόνου όσο και καθυστερήσεις (*lags*) 10 ωρών για τις βασικές μεταβλητές **DEWP** και **TEMP**, οι οποίες είχαν τη μεγαλύτερη σημασία σύμφωνα με μια αξιολόγηση σημασίας χαρακτηριστικών που έγινε. Τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες και διαγράμματα.

Αποτελέσματα

Model	MSE	MAE	RMSE	R^2
AdaBoostRegressor	8453.78	78.49	91.94	-0.09
DecisionTreeRegressor	2711.57	34.69	52.07	0.65
ElasticNet	5249.09	51.61	72.45	0.32

ExtraTreesRegressor	93.46	4.87	9.67	0.99
GaussianProcessRegressor	**0.02**	**0.00**	**0.15**	**1.00**
GradientBoostingRegressor	2965.72	37.45	54.46	0.62
KNeighborsRegressor	1779.23	25.84	42.18	0.77
Lasso	5251.30	51.61	72.47	0.32
LinearRegression	5248.28	51.63	72.45	0.32
MLPRegressor	2517.65	34.42	50.18	0.68
RandomForestRegressor	**276.43**	**10.86**	**16.63**	**0.96**
Ridge	5248.38	51.62	72.45	0.32
SVR	4498.01	40.82	67.07	0.42

Αποτελέσματα στο Σετ Εκπαίδευσης

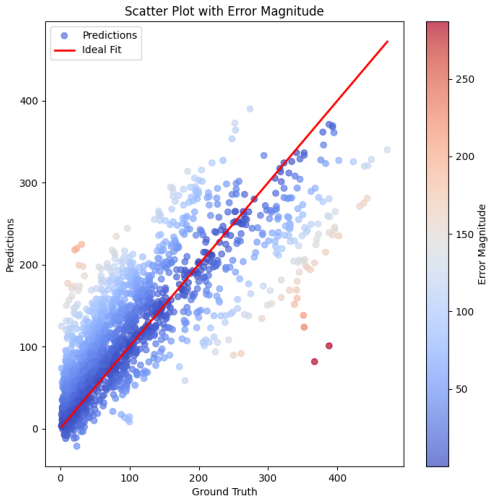
Model	**MSE**	**MAE**	**RMSE**	**R²**
AdaBoostRegressor	7801.71	74.43	88.33	-0.06
DecisionTreeRegressor	5639.23	49.08	75.09	0.23
ElasticNet	5813.72	52.36	76.25	0.21
ExtraTreesRegressor	**3785.04**	**41.89**	**61.52**	**0.49**
GaussianProcessRegressor	5871.11	53.41	76.62	0.20
GradientBoostingRegressor	4327.83	44.54	65.79	0.41
KNeighborsRegressor	4835.83	47.41	69.54	0.34
Lasso	5811.02	52.28	76.23	0.21
LinearRegression	5819.15	52.41	76.28	0.21
MLPRegressor	4226.96	45.00	65.02	0.43
RandomForestRegressor	4012.02	42.59	63.34	0.46
Ridge	5817.26	52.39	76.27	0.21
SVR	5176.07	44.10	71.94	0.30

Αποτελέσματα στο Σετ Επικύρωσης

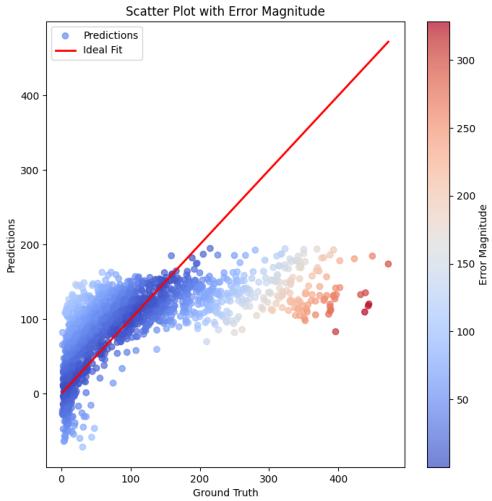
Model	**MSE**	**MAE**	**RMSE**	**R²**
AdaBoostRegressor	10456.36	89.82	102.26	-0.20
DecisionTreeRegressor	5919.35	51.59	76.94	0.32
ElasticNet	5006.06	49.23	70.75	0.43
ExtraTreesRegressor	3313.75	41.56	57.57	0.62
GaussianProcessRegressor	8152.18	57.39	90.29	0.07
GradientBoostingRegressor	3101.30	41.23	55.69	0.64
KNeighborsRegressor	4111.16	44.16	64.12	0.53
Lasso	5007.24	49.24	70.76	0.43
LinearRegression	5001.40	49.22	70.72	0.43
MLPRegressor	**3087.18**	**40.87**	**55.56**	**0.65**
RandomForestRegressor	3286.03	40.56	57.32	0.62
Ridge	5003.24	49.23	70.73	0.43
SVR	4585.28	41.77	67.71	0.47

Αποτελέσματα στο Σετ Ελέγχου

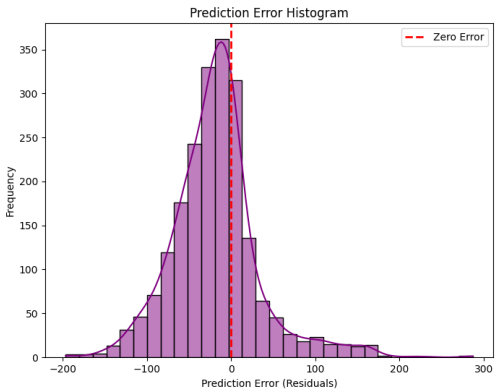
Τα αποτελέσματα δείχνουν ότι το μοντέλο **MLPRegressor** είχε την καλύτερη απόδοση στο σετ ελέγχου, ενώ το **AdaBoostRegressor** εμφάνισε τις χειρότερες επιδόσεις.



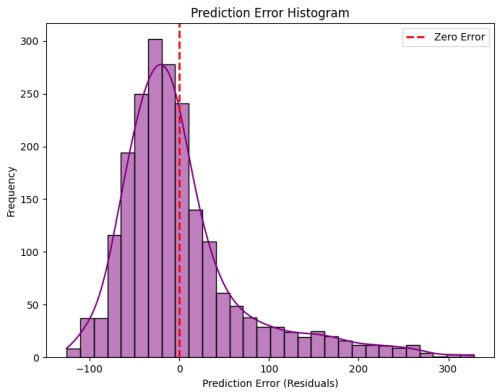
(a) **MLPRegressor:** Προβλέψεις vs Πραγματικές Τιμές



(b) **ElasticNet:** Προβλέψεις vs Πραγματικές Τιμές



(c) **MLPRegressor:** Ιστόγραμμα Σφαλμάτων



(d) **ElasticNet:** Ιστόγραμμα Σφαλμάτων

Αποτελέσματα για MLP και ElasticNet Regressors: Προβλέψεις vs Πραγματικές Τιμές (πάνω) και Ιστογράμματα Σφαλμάτων (κάτω).

Η Δυσκολία Της Εύρεσης Συνοχής Των Δεδομένων

Παρά την εκτενή ανάλυση και τη διεξαγωγή έξι διαφορετικών πειραμάτων, τα αποτελέσματα δεν έδειξαν σημαντική βελτίωση στη συνολική απόδοση των μοντέλων. Το καλύτερο αποτέλεσμα επιτεύχθηκε στο έκτο πείραμα, όπου εφαρμόστηκαν καθυστερήσεις 10 ωρών στις πιο σημαντικές μεταβλητές (DEWP και TEMP), σύμφωνα με την ανάλυση σημαντικότητας χαρακτηριστικών.
Αυτό θέτει την ανάγκη για αναθεώρηση της προσέγγισής μας, ως προς την κατανόηση της φύσης των δεδομένων.

Περαιτέρω Ανάλυση Των Δεδομένων

Για την εις βάθος διερεύνηση της συσχέτισης μεταξύ του $pm2.5$ και άλλων μεταβλητών (*dewp*, *temp*, *Iwd*, *cbwd*), εφαρμόσαμε τρεις διαφορετικές προσεγγίσεις ανάλυσης και

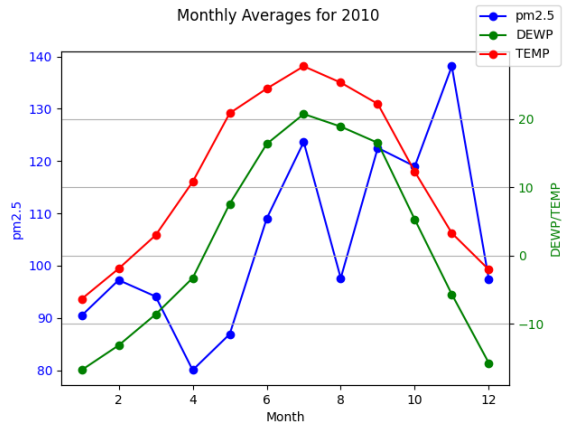
απεικόνισης των δεδομένων. Ο στόχος ήταν να εντοπίσουμε συσχετίσεις που θα μπορούσαν να αξιοποιηθούν σε μοντέλα πρόβλεψης.

Μεθοδολογία

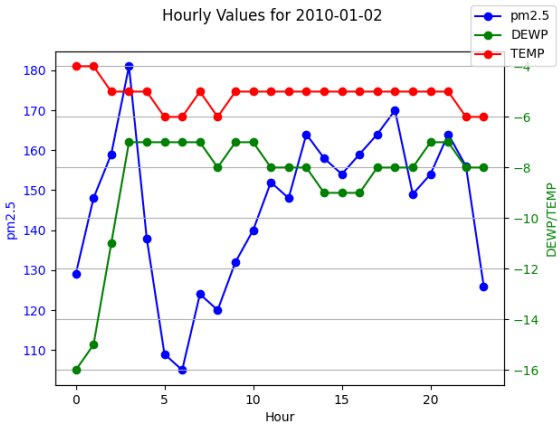
1. **Μηνιαία ανάλυση**: Υπολογίσαμε τους μηνιαίους μέσους όρους για κάθε έτος, με τις τιμές του $pm2.5$, του $dewp$ και του $temp$ να απεικονίζονται σε διαφορετικούς άξονες y.
2. **Ημερήσια ανάλυση**: Υπολογίσαμε τους ημερήσιους μέσους όρους για κάθε μήνα, εξετάζοντας τη συσχέτιση των μεταβλητών σε μεσαία χρονικά διαστήματα.
3. **Ωριαία ανάλυση**: Υπολογίσαμε τους ωριαίους μέσους όρους για κάθε ημέρα, για την παρατήρηση ταχύτερων διακυμάνσεων.

Αποτελέσματα

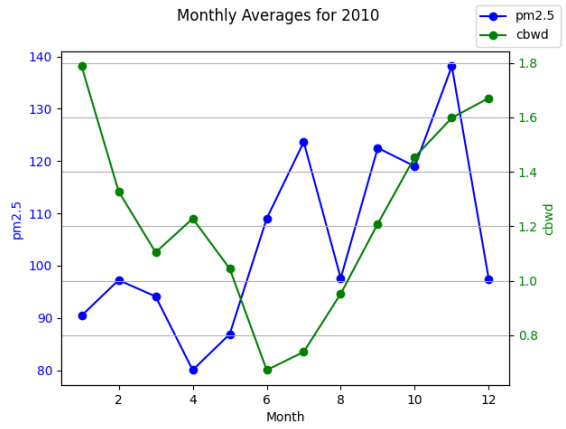
- **Μηνιαία και Ωριαία ανάλυση**: Δεν παρατηρήθηκε σημαντική συσχέτιση μεταξύ του $pm2.5$ και των μεταβλητών $DEWP$, $TEMP$, $cbwd$ και Iws, όπως φαίνεται στο Σχήμα Ανάλυσης 1. Η μηνιαία και ωριαία ανάλυση δεν ανέδειξαν κοινές τάσεις που να μπορούν να αξιοποιηθούν για προβλέψεις.
- **Ημερήσια ανάλυση**: Παρατηρήθηκε ισχυρή συσχέτιση μεταξύ του $pm2.5$ και των $DEWP$, $cbwd$ και Iws, όπως φαίνεται στο Σχήμα Ανάλυσης 2. Οι ημερήσιες τάσεις ανέδειξαν πιο ξεκάθαρα μοτίβα, καθιστώντας αυτήν την ανάλυση πιο χρήσιμη για μοντέλα πρόβλεψης.



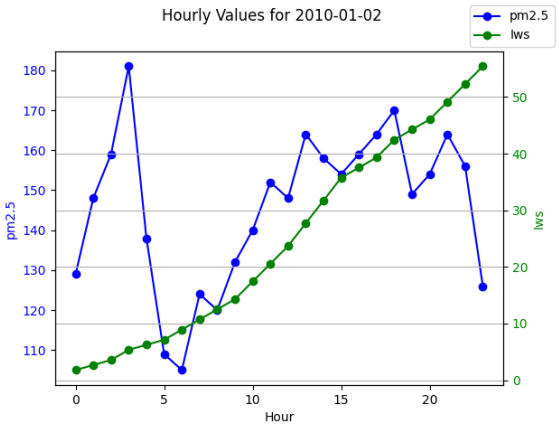
(a) Μηνιαία ανάλυση: $DEWP$, $TEMP$, $pm2.5$



(b) Ωριαία ανάλυση: $DEWP$, $TEMP$, $pm2.5$

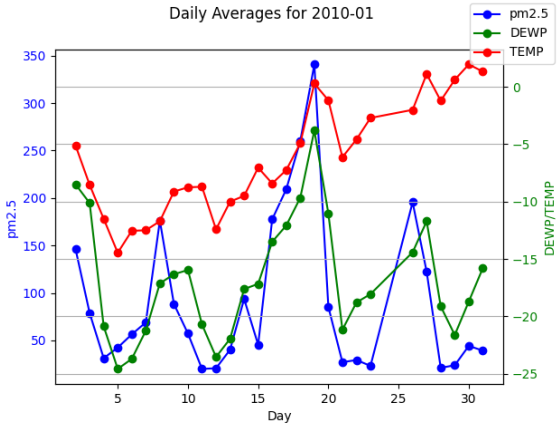


(c) Μηνιαία ανάλυση: $cbwd$, $pm2.5$

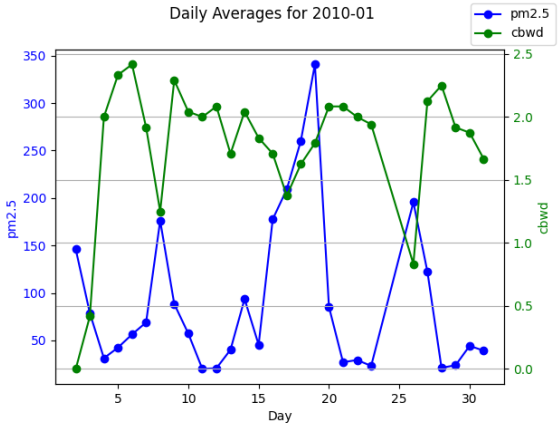


(d) Ωριαία ανάλυση: Iws, $pm2.5$

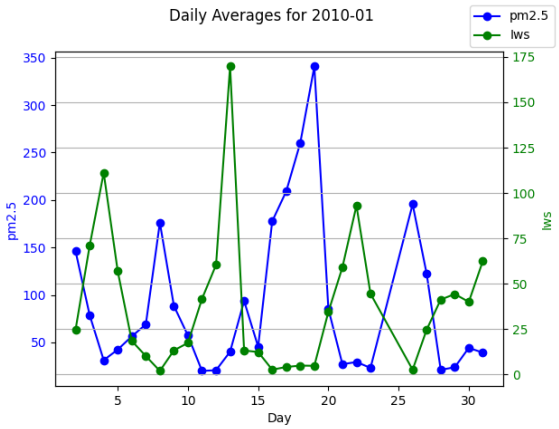
Σχήμα Ανάλυσης 1: Τα παραπάνω γραφήματα υποδεικνύουν την αδύναμη συσχέτιση μεταξύ $pm2.5$ και άλλων μεταβλητών στις μηνιαίες και ωριαίες αναλύσεις.



(a) Ημερήσια ανάλυση: $dewp$, $temp$, $pm2.5$



(b) Ημερήσια ανάλυση: $cbwd$, $pm2.5$



(c) Ημερήσια ανάλυση: Iws, $pm2.5$

Σχήμα Ανάλυσης 2: Τα παραπάνω γραφήματα δείχνουν την ισχυρή συσχέτιση μεταξύ $pm2.5$ και των $dewp$, $cbwd$ και Iws στις ημερήσιες αναλύσεις.

Σχόλια:

- Στο Σχήμα Ανάλυσης 2(a), η πράσινη γραμμή ($dewp$) ακολουθεί σχεδόν πιστά τις διακυμάνσεις του $pm2.5$, κάτι που υποδεικνύει ισχυρή συσχέτιση και σημασία της μεταβλητής για την πρόβλεψη του $pm2.5$.
- Στο Σχήμα Ανάλυσης 2(b), παρατηρούνται αντίστροφες τάσεις μεταξύ του $cbwd$ και του $pm2.5$, δηλαδή όταν το ένα αυξάνεται, το άλλο μειώνεται. Αυτή η σταθερή αντίστροφη συσχέτιση είναι αξιοσημείωτη.
- Στο Σχήμα Ανάλυσης 2(c), το Iws παρουσιάζει ισχυρή συσχέτιση με το $pm2.5$, με μία μικρή χρονική καθυστέρηση.

Μετατροπή Σε Πρόβλημα Κατηγοριοποίησης

Η ανάλυση που πραγματοποιήθηκε κατέδειξε τις περιορισμένες δυνατότητες ακριβούς πρόβλεψης των τιμών του $pm2.5$ μέσω μοντέλων παλινδρόμησης. Παρά τις διάφορες προσεγγίσεις και επεξεργασίες των δεδομένων, τα αποτελέσματα υπήρξαν απογοητευτικά, με χαμηλές τιμές R^2 και υψηλά σφάλματα. Αυτή η αδυναμία υποδεικνύει ότι το πρόβλημα, όπως ορίστηκε, ενδέχεται να είναι ιδιαίτερα περίπλοκο για τις υπάρχουσες παραδοχές και τα χρησιμοποιούμενα μοντέλα.

Νέα Προσέγγιση: Binary Classification Προτείνεται η μετατροπή του προβλήματος από παλινδρόμηση (regression) σε δυαδική κατηγοριοποίηση (binary classification). Αντί να προβλέπεται η ακριβής τιμή του $pm2.5$, η νέα προσέγγιση εστιάζει στην πρόβλεψη της κατεύθυνσης μεταβολής της τιμής ($\uparrow$ ή $\downarrow$) σε σχέση με την τρέχουσα τιμή. Η προσέγγιση αυτή βασίζεται στις παρακάτω παραδοχές και πλεονεκτήματα:

- **Απλούστερη μοντελοποίηση**: Η δυαδική κατηγοριοποίηση μειώνει την πολυπλοκότητα του προβλήματος, εστιάζοντας στην πρόβλεψη τάσεων αντί ακριβών αριθμητικών τιμών. Με αυτόν τον τρόπο, τα μοντέλα μπορούν να αξιοποιήσουν πιο εύκολα τις υπάρχουσες σχέσεις μεταξύ των χαρακτηριστικών.
- **Ευθυγράμμιση με τις συσχετίσεις**: Οι συσχετίσεις που παρατηρήθηκαν στην ημερήσια ανάλυση, ιδιαίτερα για μεταβλητές όπως το $dewp$ και το Iws, υποδεικνύουν ότι είναι πιο εφικτή η εκτίμηση της κατεύθυνσης μεταβολής του $pm2.5$ σε σχέση με την ακριβή τιμή του.
- **Πρακτική χρησιμότητα**: Σε πολλά σενάρια περιβαλλοντικής παρακολούθησης, η κατεύθυνση της μεταβολής είναι πιο χρήσιμη από την ακριβή πρόβλεψη. Για παράδειγμα, σε εφαρμογές έγκαιρης προειδοποίησης, είναι πιο σημαντικό να γνωρίζουμε αν οι συνθήκες επιδεινώνονται, ώστε να ληφθούν έγκαιρα μέτρα.

Ημερήσια Ανάλυση έναντι Ωριαίας Η επιλογή ημερήσιας ανάλυσης, αντί ωριαίας, δικαιολογείται από τα ευρήματα της αρχικής μελέτης. Όπως καταδείχθηκε, οι ημερήσιες τάσεις παρουσιάζουν πιο σταθερές και ισχυρές συσχετίσεις μεταξύ του $pm2.5$ και άλλων χαρακτηριστικών, όπως το $dewp$, το $cbwd$ και το Iws. Αντίθετα, οι ωριαίες διακυμάνσεις χαρακτηρίζονται από αυξημένο θόρυβο, καθιστώντας την κατηγοριοποίηση πιο επισφαλή. Συνεπώς, η ημερήσια προσέγγιση όχι μόνο ευθυγραμμίζεται καλύτερα με τις παρατηρήσεις μας, αλλά και καθιστά τη μετατροπή του προβλήματος σε binary classification πιο εφαρμόσιμη.

Αποτελέσματα Μοντέλων Κατηγοριοποίησης

Μετά τη μετατροπή του προβλήματος από παλινδρόμηση σε δυαδική κατηγοριοποίηση, τα αποτελέσματα των μοντέλων έδειξαν σαφή βελτίωση όσον αφορά την απόδοση. Η ανάλυση των $pm2.5\_change$ (αύξηση ή μείωση) οδήγησε σε υψηλές τιμές ακρίβειας, ανάκλησης και $F1$-score τόσο στο σετ επικύρωσης όσο και στο σετ ελέγχου.

Εδώ, κάναμε συνολικά δύο παρόμοια πειράματα. Στο πρώτο πείραμα χρησιμοποιήθηκαν καθυστερήσεις (lags) 1 και 2 ημερών για όλες τις μεταβλητές, εκτός από το στόχο (PM2.5), ενώ στο δεύτερο πείραμα χρησιμοποιήθηκαν καθυστερήσεις απο 1 εώς 10 ημέρες.

Ιδιαίτερη εντύπωση προκάλεσε η απόδοση των μοντέλων Logistic Regression, Random Forest και SVM, τα οποία εμφάνισαν τις καλύτερες τιμές $F1$-score στο σετ ελέγχου και στα δύο πειράματα.
Παρακάτω παρουσιάζονται τα αποτελέσματα των καλύτερων μοντέλων, με έμφαση στις τιμές ακρίβειας ($accuracy$) και $F1$-score για το σετ επικύρωσης και το σετ ελέγχου.

Πείραμα 1: Καθυστερήσεις 1 και 2 Ημερών

Αποτελέσματα Επικύρωσης

Μοντέλο	Precision	Recall	Weighted F1-Score
Random Forest	0.79	0.79	0.7868
Gradient Boosting	0.78	0.78	0.7774
Logistic Regression	**0.82**	**0.82**	**0.8232**
Decision Tree	0.74	0.74	0.7387
K-Nearest Neighbors	0.75	0.74	0.7340
SVM	**0.83**	**0.83**	**0.8340**

Πίνακας 1: Αποτελέσματα επικύρωσης για τα διάφορα μοντέλα κατηγοριοποίησης του πειράματος 1.

Αποτελέσματα Ελέγχου

Μοντέλο	Precision	Recall	Weighted F1-Score
Random Forest	0.83	0.82	0.8094
Gradient Boosting	0.80	0.79	0.7900
Logistic Regression	**0.85**	**0.85**	**0.8457**
Decision Tree	0.79	0.79	0.7869
K-Nearest Neighbors	0.80	0.79	0.7900
SVM	0.85	0.84	0.8337

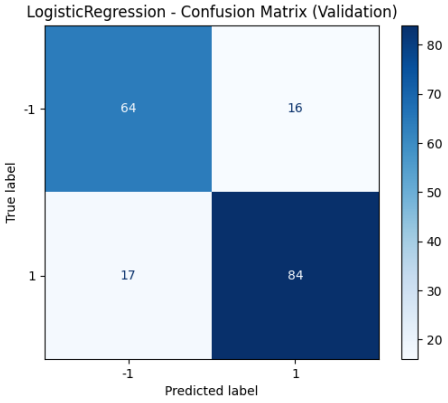
Πίνακας 2: Αποτελέσματα ελέγχου για τα διάφορα μοντέλα κατηγοριοποίησης του πειράματος 1.

Τα αποτελέσματα δείχνουν ότι η μετατροπή σε πρόβλημα κατηγοριοποίησης βελτίωσε σημαντικά την απόδοση των μοντέλων. Συγκεκριμένα:

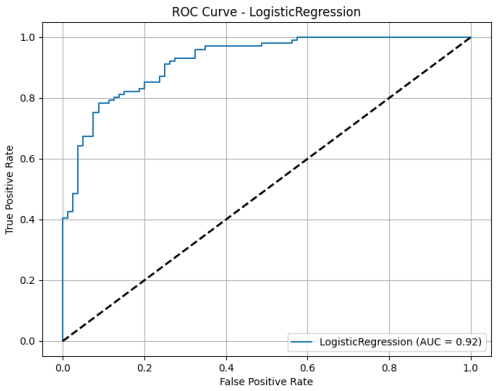
- Το Logistic Regression εμφάνισε την καλύτερη απόδοση συνολικά, με $WeightedF1-Score$ ίσο με 0.8457 στο σετ ελέγχου.
- Το SVM και το Random Forest κατέγραψαν επίσης υψηλές τιμές $F1$-score, δείχνοντας τη δυνατότητά τους να μοντελοποιούν τις αλλαγές του $pm2.5$.

- Τα μοντέλα Gradient Boosting και K-Nearest Neighbors είχαν παρόμοιες επιδόσεις, αλλά χαμηλότερες συγκριτικά με τα Logistic Regression και SVM.
- Το Decision Tree παρουσίασε ικανοποιητική απόδοση, αλλά ήταν εμφανώς χαμηλότερο από τα υπόλοιπα μοντέλα.

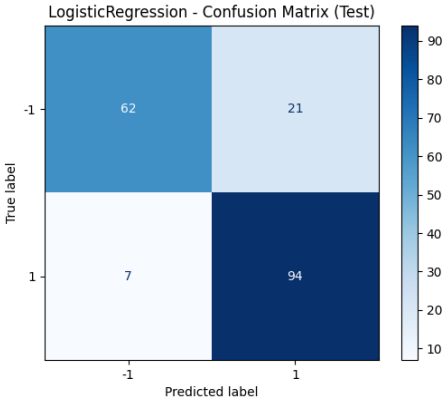
Η νέα προσέγγιση κατέδειξε ότι η πρόβλεψη της μεταβολής του $pm2.5$ είναι πιο ρεαλιστική και αποδοτική, ιδιαίτερα όταν χρησιμοποιούνται τα σωστά χαρακτηριστικά και ακολουθούνται κατάλληλες τεχνικές κατηγοριοποίησης.
Για να αξιολογήσουμε περαιτέρω την απόδοση των μοντέλων μας, παρουσιάζουμε παρακάτω τα διαγράμματα ROC Curve και Confusion Matrix για τα δύο καλύτερα μοντέλα, Logistic Regression και SVM. Αυτά τα διαγράμματα παρέχουν μια πιο οπτική κατανόηση της απόδοσης των μοντέλων τόσο στο σετ επικύρωσης όσο και στο σετ ελέγχου.



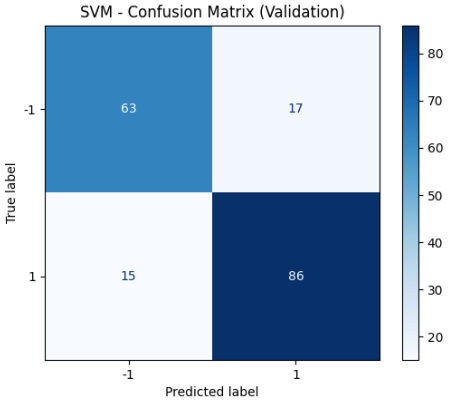
(a) Logistic Regression: Confusion Matrix (Validation)



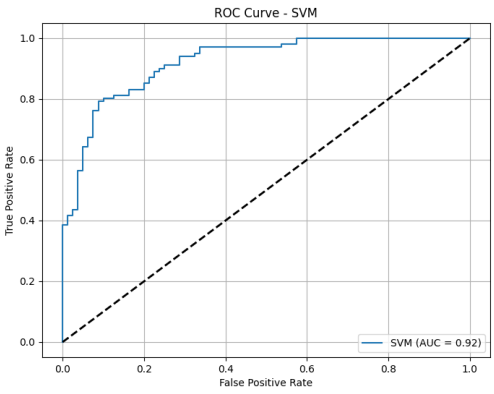
(b) Logistic Regression: ROC Curve (Test)



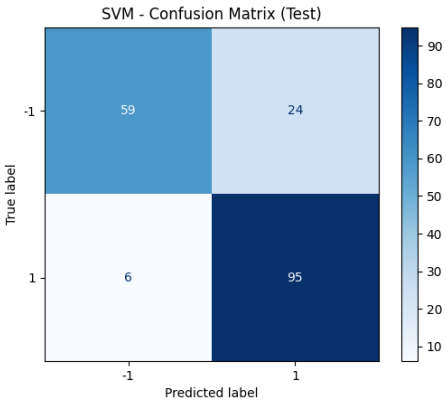
(c) Logistic Regression: Confusion Matrix (Test)



(d) SVM: Confusion Matrix (Validation)



(e) SVM: ROC Curve (Test)



(f) SVM: Confusion Matrix (Test)

Συγκεντρωτική απεικόνιση Confusion Matrices και ROC Curves για Logistic Regression και SVM.

Τα διαγράμματα ROC επιβεβαιώνουν την υψηλή απόδοση των Logistic Regression και SVM, με τις καμπύλες να πλησιάζουν την πάνω αριστερή γωνία, υποδεικνύοντας υψηλή ευαισθησία και ειδικότητα. Τα Confusion Matrices αποδεικνύουν την ικανότητα των μοντέλων να ταξινομούν σωστά τις κλάσεις, με λίγα λάθη σε κάθε σετ δεδομένων.

Πείραμα 2: Καθυστερήσεις από 1 εώς και 10 Ημερών

Αποτελέσματα Επικύρωσης

Μοντέλο	Precision	Recall	Weighted F1-Score
Random Forest	0.75	0.75	0.7479
Gradient Boosting	0.78	0.78	0.7774
Logistic Regression	**0.83**	**0.83**	**0.8290**
Decision Tree	0.72	0.72	0.7171
K-Nearest Neighbors	0.65	0.65	0.6490
SVM	0.86	0.86	0.8618

Πίνακας 3: Αποτελέσματα επικύρωσης για τα διάφορα μοντέλα κατηγοριοποίησης του πειράματος 2.

Αποτελέσματα Ελέγχου

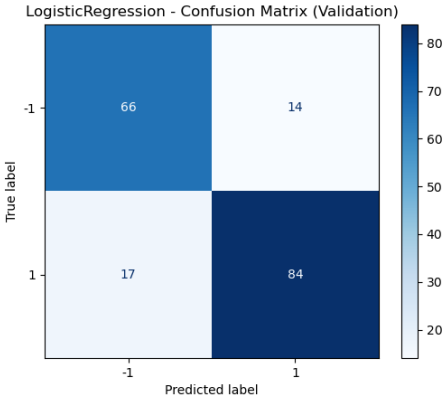
Μοντέλο	Precision	Recall	Weighted F1-Score
Random Forest	0.81	0.80	0.7946
Gradient Boosting	0.81	0.80	0.8005
Logistic Regression	**0.86**	**0.86**	**0.8575**
Decision Tree	0.69	0.68	0.6795
K-Nearest Neighbors	0.71	0.71	0.6973
SVM	0.85	0.85	0.8462

Πίνακας 4: Αποτελέσματα ελέγχου για τα διάφορα μοντέλα κατηγοριοποίησης του πειράματος 2.

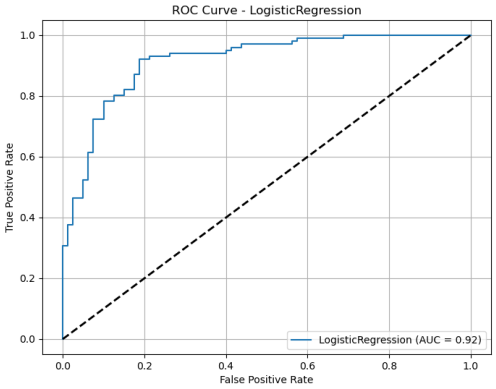
Συμπερασματικά, αυτό που μπορούμε να παρατηρήσουμε σε σύγκριση με τα αποτελέσματα του πειραμάταος 2 είναι:

- Στα περισσότερα μοντέλα παρατηρούμε πολύ κοντινά αποτελέσματα
- Στα μοντέλα Decision Tree και K-Nearest Neighbors παρατηρούμε μία αρκετά μεγάλη πτώση.

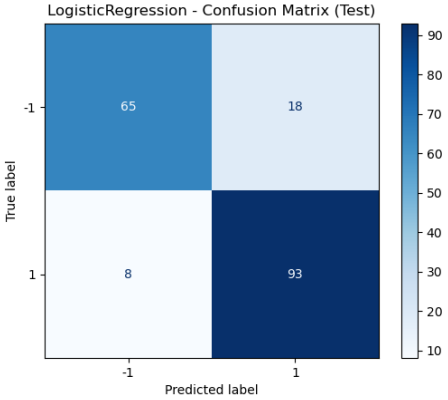
Η υποβάθμιση της απόδοσης των K-Nearest Neighbors (KNN) και των δέντρων απόφασης όταν ενσωματώνονται οι τιμές PM2.5 από τις τελευταίες 10 ημέρες μπορεί να αποδοθεί σε διάφορους παράγοντες. Πρώτον, ο KNN υποφέρει από την κατάρα της διαστατικότητας, καθώς η αύξηση του αριθμού των χαρακτηριστικών διευρύνει το χώρο των χαρακτηριστικών, καθιστώντας τους υπολογισμούς απόστασης λιγότερο σημαντικούς. Καθώς αυξάνεται η διάσταση, η ευκλείδεια απόσταση μεταξύ των σημείων δεδομένων γίνεται λιγότερο διακριτική, οδηγώντας σε μειωμένη ακρίβεια ταξινόμησης. Δεύτερον, τα Δένδρα Απόφασης είναι επιρρεπή σε υπερπροσαρμογή όταν εκτίθενται σε μεγαλύτερο αριθμό χαρακτηριστικών, ιδίως όταν αυτά τα χαρακτηριστικά είναι σε μεγάλο βαθμό συσχετισμένα, όπως συμβαίνει συνήθως σε δεδομένα χρονοσειρών. Με τη συμπερίληψη περισσότερων παρελθουσών τιμών PM2.5, το δέντρο γίνεται πιο πολύπλοκο, συλλαμβάνοντας θόρυβο αντί για ουσιώδη μοτίβα, γεγονός που μειώνει την ικανότητά του να γενικεύει σε αόρατα δεδομένα. Τέλος, αυτή η υποβάθμιση των επιδόσεων μπορεί επίσης να εξηγηθεί μέσω του συμβιβασμού μεροληψίας-διακύμανσης. Με λιγότερα χαρακτηριστικά (τελευταίες 2 ημέρες), και τα δύο μοντέλα είχαν μια πιο βέλτιστη ισορροπία μεταξύ προκατάληψης και διακύμανσης. Ωστόσο, με τη συμπερίληψη πρόσθετων ιστορικών δεδομένων, το KNN παρουσίασε αυξημένη διακύμανση λόγω του χώρου υψηλών διαστάσεων, ενώ τα δέντρα απόφασης προσαρμόστηκαν υπερβολικά στα δεδομένα εκπαίδευσης, οδηγώντας σε μειωμένη απόδοση ταξινόμησης.



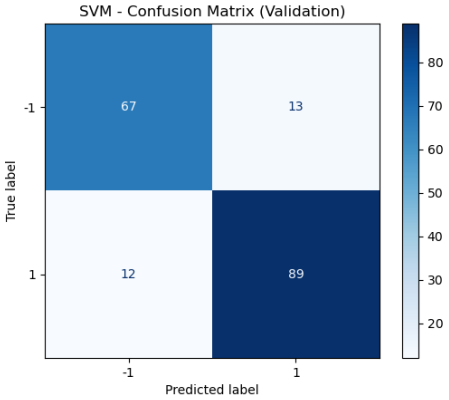
(a) Logistic Regression: Confusion Matrix (Validation)



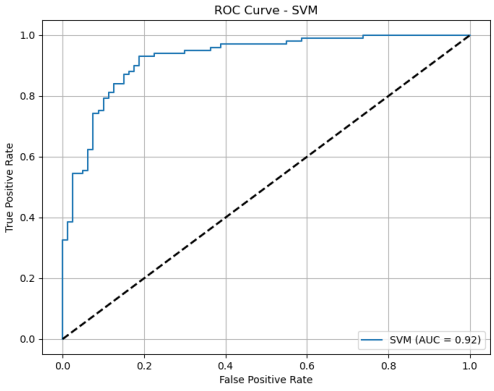
(b) Logistic Regression: ROC Curve (Test)



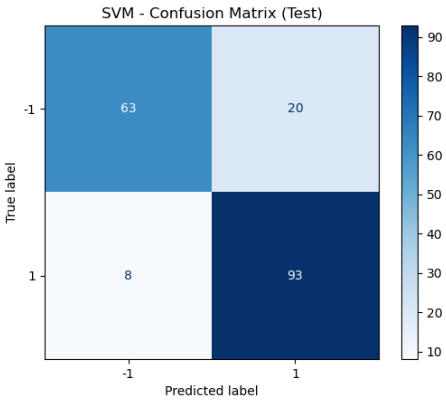
(c) Logistic Regression: Confusion Matrix (Test)



(d) SVM: Confusion Matrix (Validation)



(e) SVM: ROC Curve (Test)



(f) SVM: Confusion Matrix (Test)

Συγκεντρωτική απεικόνιση Confusion Matrices και ROC Curves για Logistic Regression και SVM.

Όπως και πριν, έτσι και τώρα τα διαγράμματα ROC επιβεβαιώνουν την υψηλή απόδοση των Logistic Regression και SVM, με τις καμπύλες να πλησιάζουν την πάνω αριστερή γωνία

Σύναψη

Στη μελέτη αυτή, εξετάσαμε τη δυνατότητα πρόβλεψης της συγκέντρωσης $pm2.5$ στην ατμόσφαιρα του Πεκίνου μέσω τεχνικών μηχανικής μάθησης. Ξεκινήσαμε με την ανάλυση και επεξεργασία ενός συνόλου δεδομένων που περιλαμβάνει ωριαίες μετρήσεις μετεωρολογικών μεταβλητών, όπως $dewp$, $temp$, $pres$, $cbwd$, Iws, Ir κ.α. Το αρχικό πρόβλημα αντιμετωπίστηκε ως πρόβλημα παλινδρόμησης, αλλά λόγω των χαμηλών επιδόσεων των μοντέλων, έγινε μετατροπή του σε δυαδικό πρόβλημα κατηγοριοποίησης, όπου προβλέπουμε τη μεταβολή (↑ ή ↓) της τιμής του $pm2.5$.
Κατά τη διάρκεια της ανάλυσης:

- **Παλινδρόμηση:** Δοκιμάσαμε διάφορα μοντέλα, όπως Random Forest, Gradient Boosting, MLP Regressors κ.α. χρησιμοποιώντας διαφορετικές τεχνικές κατασκευής χαρακτηριστικών, όπως καθυστερήσεις (lags) και κυκλική κωδικοποίηση χρόνου. Τα αποτελέσματα έδειξαν ότι ακόμα και τα καλύτερα μοντέλα είχαν περιορισμένες επιδόσεις, με R^2 να φτάνει μέχρι 0.65 για τα δεδομένα ελέγχου, καταδεικνύοντας την πολυπλοκότητα του προβλήματος.

- **Ανάλυση Δεδομένων:** Μελετήσαμε τη συνοχή των μεταβλητών και τις συσχετίσεις τους με το $pm2.5$ σε διαφορετικά χρονικά διαστήματα (ωριαία, ημερήσια, μηνιαία). Η ανάλυση έδειξε ότι οι ημερήσιες μέσες τιμές παρέχουν μεγαλύτερη συνοχή και σταθερότητα σε σχέση με τις ωριαίες, οι οποίες εμφανίζουν έντονο θόρυβο. Ειδικότερα, οι μεταβλητές $dewp$, $cbwd$ και Iws παρουσίασαν ισχυρή συσχέτιση με τις διακυμάνσεις του $pm2.5$, καθιστώντας τες σημαντικές για την πρόβλεψη της μεταβολής του.

- **Κατηγοριοποίηση:** Μετατρέψαμε το πρόβλημα σε δυαδικό πρόβλημα κατηγοριοποίησης, εστιάζοντας στη μεταβολή του $pm2.5$ ($pm2.5\_change$). Δοκιμάσαμε μοντέλα όπως Logistic Regression, SVM, Random Forest κ.α.. Τα αποτελέσματα έδειξαν σαφή βελτίωση, με το Logistic Regression να καταγράφει την καλύτερη απόδοση, πετυχαίνοντας $F1$-score 0.8575 στο σετ ελέγχου. Παράλληλα, τα SVM και Random Forest σημείωσαν εξαιρετικά αποτελέσματα, δείχνοντας την ικανότητά τους να αξιοποιούν τα διαθέσιμα χαρακτηριστικά.

Παρουσιάσαμε αναλυτικά τα αποτελέσματα μέσω πινάκων που συνοψίζουν την απόδοση των μοντέλων, καθώς και διαγραμμάτων, όπως Confusion Matrices και ROC Curves, που αναδεικνύουν την ισορροπία ευαισθησίας και ειδικότητας.
Η μελέτη μας κατέδειξε τη δυσκολία ακριβούς πρόβλεψης της συγκέντρωσης $pm2.5$ μέσω παλινδρόμησης, ενώ η μετατροπή σε δυαδικό πρόβλημα κατηγοριοποίησης προσέφερε σαφή πλεονεκτήματα. Τα αποτελέσματα υπογραμμίζουν τη σημασία των μεταβλητών $dewp$, $cbwd$ και Iws για την πρόβλεψη των μεταβολών του $pm2.5$, ενώ δείχνουν ότι η κατηγοριοποίηση με σωστή κατασκευή χαρακτηριστικών μπορεί να παρέχει πιο ρεαλιστικές και χρήσιμες προβλέψεις.

Πηγές Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι διαθέσιμο στη διεύθυνση: `Beijing Air Quality`