ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΟ ΠΛΗΡΟΦΟΡΙΑΚΟ ΣΥΣΤΗΜΑ ΑΝΙΧΝΕΥΣΗΣ ΕΥΑΙΣΘΗΤΩΝ ΠΡΟΣΩΠΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ

ΑΙΜΙΛΙΟΣ ΒΑΤΙΚΙΩΤΗΣ ΡΟΚΚΙ

ΑΜ: 2114037

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Σταμούλης Γεώργιος

…….Βαθμίδα…….

ΣΥΝΕΠΙΒΛΕΠΩΝ

Σπαθούλας Γεώργιος

…….Βαθμίδα…….

Λαμία …… έτος ……

Αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

Αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

Αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας

Αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας

1. Εισαγωγή

Στο τέλος της δεκαετίας που διανύουμε, η ανθρωπότητα έχει ενταχθεί, σε τεράστιο βαθμό πλέον, στην εποχή της τεχνολογίας, που εξακολουθεί να αναπτύσσεται ραγδαία σε όλους τους τομείς της ανθρώπινης δραστηριότητας. Όλοι μας ανεξαιρέτως καταφεύγουμε καθημερινά στον ηλεκτρονικό μας υπολογιστή, με εξατομικευμένες διαφοροποιήσεις τόσο όσον αφορά τη διάρκεια όσο και το λόγο της ενασχόλησης το κινητό είναι αχώριστο και απολύτως απαραίτητο εργαλείο επικοινωνίας, όπως και τα email και τα μέσα κοινωνικής δικτύωσης, σε προσωπικό και επαγγελματικό επίπεδο.

Οι εξελίξεις της τεχνολογίας, ιδιαίτερα του κλάδου της Πληροφορικής, προτρέπουν τις επιχειρήσεις να αναζητούν νέες μεθόδους, σύγχρονες εσωτερικές και εξωτερικές στρατηγικές που στοχεύουν στη βελτιστοποίηση των συνθηκών εργασίας, στην αξιοποίηση του ανθρώπινου δυναμικού και στην αναζήτηση πηγών κέρδους, μέσα από το Στοχευμένο Μάρκετινγκ (Target Marketing), την Επιχειρηματική Ευφυΐα (Business Intelligence) και την ανάλυση των πληροφοριών (data analysis) που ήδη διατίθενται και συγκεντρώνονται σε καθημερινή βάση.

Σε κάθε περίπτωση, τα οφέλη της τεχνολογίας είναι πρόδηλα και αδιαμφισβήτητα. Ωστόσο, συνάμα με τις θετικές πτυχές της, τα τελευταία χρόνια αναδύθηκαν φαινόμενα κατάχρησης αυτής της τεχνολογίας, που αποσκοπούν στην επίτευξη είτε προσωπικής ικανοποίησης είτε επαγγελματικών οφελών, παραγκωνίζοντας εντελώς τις βασικές αρχές των ανθρωπίνων δικαιωμάτων και τα πνευματικά δικαιώματα συνυφασμένα ενδεχομένως με τον ίδιο τον σκοπό κάθε συγκεκριμένης εργασίας.

Για την επίλυση των προεκτεθέντων ζητημάτων, η Ευρωπαϊκή Ένωση, κατά τη συνεδρίαση του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου που πραγματοποιήθηκε τον Απρίλιο του 2016, προχώρησε στην εκπόνηση νομοθετικής πράξης του Κανονισμού **"***Για την προστασία των φυσικών προσώπων έναντι της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα και την ελεύθερη κυκλοφορία των δεδομένων αυτών (GDPR[1])***"** καθώς και, ταυτόχρονα, στην κατάργηση της οδηγίας 95/46/*ΕΚ (Γενικός Κανονισμός για την Προστασία Δεδομένων).*

Ο κανονισμός GDPR εφαρμόζεται, εν όλω ή εν μέρει, στην αυτοματοποιημένη επεξεργασία δεδομένων προσωπικού χαρακτήρα καθώς και στη μη αυτοματοποιημένη επεξεργασία των δεδομένων αυτών, τα οποία περιλαμβάνονται ήδη, ή πρόκειται να περιληφθούν, σε σύστημα αρχειοθέτησης.

---

[1] Με τη συντομογραφία GDPR – ΓΚΠΔ θα απευθυνόμαστε από δω και στο εξής για το Γενικό Κανονισμό Προστασίας Δεδομένων και General Data Protection Regulation αντίστοιχα.

Συγκεκριμένα, κάθε είδος ή τύπος δεδομένων τα οποία επεξεργάζονται από οποιοδήποτε φορέα, οργανισμό ή επιχείρηση, πρέπει να μετασχηματίζονται κατάλληλα ώστε να συμβαδίζουν με τις οδηγίες που θεσπίζονται στον Κανονισμό, πριν από την έμπρακτη υιοθέτησή τους. Η επεξεργασία και η αξιοποίηση των ευαίσθητων δεδομένων επιτρέπονται αποκλειστικά υπό την προϋπόθεση να μην υφίσταται ουδεμία δυνατότητα συσχέτισής τους με κάποιο συγκεκριμένο φυσικό πρόσωπο.

Ως επακόλουθο της υιοθέτησης της εν λόγω κοινοτικής οδηγίας προβάλλεται η ανάγκη προσαρμογής των λογισμικών εφαρμογών του βιομηχανικού και επιχειρηματικού τομέα, μέσα από λίαν πολύπλοκες και δυσχερείς διεργασίες λαμβανομένου υπόψη του μεγάλου όγκου των δεδομένων και των πληροφοριών, καθώς όλες οι συναλλαγές, η επικοινωνία, η κάλυψη αναγκών σε προσωπικό και επαγγελματικό επίπεδο και γενικότερα η ανταλλαγή πληροφοριών πραγματοποιούνται πλέον μέσω του διαδικτύου και των πληροφοριακών συστημάτων.

Χάρη στις επιστημονικές και τεχνικές γνώσεις που διατίθενται από τον κλάδο της πληροφορικής[2], αναπτύσσονται εφαρμογές, λύνονται προβλήματα επικοινωνίας, καλύπτονται ανάγκες διαχείρισης και υλοποίησης έργων και παράλληλα συσσωρεύεται άπειρος αριθμός πληροφοριών και δεδομένων σε συστήματα αρχειοθέτησης, στις οποίες έχει πρόσβαση ο εκάστοτε οργανισμός ή φορέας, δημόσιος και ιδιωτικός.

Επιπλέον, πρέπει να επισημανθεί ότι κανείς δεν μπορεί να εγγυηθεί την εξάλειψη μελλοντικών προβλημάτων μετά από τη σημερινή προσαρμογή και διαμόρφωση, βάσει του Κανονισμού, των υφιστάμενων πληροφοριών.

Η εμβέλεια του υπάρχοντος όγκου των δεδομένων που ανταλλάσσεται και επεξεργάζεται από την κάθε επιχείρηση ή οργανισμό/φορέα είναι τόσο μεγάλη που κρίνεται αδύνατη η διαχείριση και η προσαρμογή τους στον υπό εξέταση νέο κανονισμό.

Νέοι και συνεχώς αναπτυσσόμενοι κλάδοι της Πληροφορικής, που συνδράμουν στην εξελικτική πορεία της τεχνολογίας, είναι η Τεχνητή Νοημοσύνη (Artificial Intelligence) και ιδιαίτερα το κομμάτι της Μηχανικής Μάθησης[3] (Machine Learning) όπως επίσης και η Εξόρυξη Δεδομένων (Data Mining).

---

[2] Ως επιστημονικές γνώσεις που προέρχονται από τον κλάδο της Πληροφορικής και θεωρούνται αναπτυσσόμενοι εννοούμε τη Τεχνητή Νοημοσύνη, τη Μηχανική Μάθηση και την Εξόρυξη Δεδομένων που ως κοινό παρανομαστή έχουν τη συλλογή και την επεξεργασία δεδομένων.

[3] Στην εργασία αυτή θα αναλύσουμε τις έννοιες και τις τεχνικές της Μηχανικής Μάθησης για την υλοποίηση του αλγορίθμου εποπτευόμενης μάθησης για την αυτόματη ανίχνευση προσωπικών πληροφοριών.

Έχει παγιοποιηθεί η απόλυτη αναγκαιότητα της κατάρτισης των παραπάνω επιστημονικών γνώσεων και εννοιών, ώστε να αποτελεί πρωταρχικός στόχος κάθε επιχείρησης η υιοθέτησή τους στις προσφερόμενες υπηρεσίες και στα διαθέσιμα προϊόντα της. Θα μπορούσε κανείς να ισχυρισθεί ότι με την αξιοποίησή τους γίνεται εφικτή η αντιμετώπιση πολλών προβλημάτων της καθημερινότητας για την κάλυψη σύγχρονων αναγκών.

Από την παραπάνω διαπίστωση, ανάγεται το συμπέρασμα ότι αποτελεί πρώτη προτεραιότητα η προσαρμογή στη νέα νομοθετική οδηγία της Ευρωπαϊκής Ένωσης. Ως εκ τούτου, κρίνεται αναγκαία η ανάπτυξη ενός «έξυπνου λογισμικού[4] που θα αναγνωρίζει τα ευαίσθητα προσωπικά δεδομένα, όπως θεσπίζει ο νέος αυτός Κανονισμός, επιλύοντας έτσι ένα πρόβλημα που, όσο αυξάνεται ο όγκος των πληροφοριών με εκθετικούς ρυθμούς τόσο πιο δύσκολη θα είναι και η αντιμετώπισή του.

Ως ευαίσθητα δεδομένα η οδηγία ορίζει:

- Δεδομένα που αφορούν την υγεία
- Βιομετρικά δεδομένα
- Προσωπικά δεδομένα
- Δεδομένα ταυτοποίησης φυσικού προσώπου

Για αυτούς τους λόγους, η εργασία αυτή εστιάζεται στον κανονισμό GDPR, αναλύει τις οδηγίες και τις απαιτήσεις του κανονισμού και, τέλος, γίνεται προσπάθεια αντιμετώπισης αυτού του ζητήματος μέσω της ανάπτυξης και υλοποίησης αυτοματοποιημένου πληροφοριακού συστήματος για την ανίχνευση ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας, με στόχο την εύρεση λύσης σε ένα σύγχρονο και επίκαιρο πρόβλημα.

[4] Με τον όρο «έξυπνο λογισμικό» εννοούμε ότι ο αλγόριθμος που θα κατασκευάσουμε εφόσον εκπαιδευτεί σωστά μέσα από τις τεχνικές Μηχανικής Μάθησης και συγκεκριμένα με τη χρήση της Γραμμικής Οπισθοδρόμησης που θα αναλύσουμε στη συνέχεια, θα είναι σε θέση να ανιχνεύει αυτόματα ευαίσθητες προσωπικές πληροφορίες.

2. Κανονισμός GDPR

2.1 Περιεχόμενα κανονισμού, επηρεαζόμενοι φορείς και υποκείμενα

Το Ευρωπαϊκό Κοινοβούλιο προχώρησε, στη συνεδρίαση του Συμβουλίου που πραγματοποιήθηκε στις 27 Απριλίου 2016, σε νομοθετική πράξη θέσπισης του κανονισμού για την προστασία των φυσικών προσώπων, καθορίζοντας τα πλαίσια της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα και της ελεύθερης κυκλοφορίας των δεδομένων αυτών. Ταυτόχρονα, καταργείται η οδηγία 95/46/ΕΚ (Γενικός Κανονισμός για την Προστασία Δεδομένων).

Οι λόγοι για τους οποίους νομοθετήθηκε αυτός ο κανονισμός είναι εξής:

- Κάθε φυσικό πρόσωπο έχει δικαίωμα στην προστασία των δεδομένων προσωπικού χαρακτήρα που το αφορούν
- Οι αρχές και οι κανόνες που διέπουν την επεξεργασία των ευαίσθητων δεδομένων θα πρέπει να σέβονται τα θεμελιώδη δικαιώματα του φυσικού προσώπου
- Η διαφύλαξη της εχεμύθειας της ιδιωτικής και οικογενειακής ζωής, της ελευθερία της σκέψης, της συνείδησης, της έκφρασης και της πληροφόρησης, καθώς επίσης της ελευθερίας του επιχειρείν.

Λόγω του διευρυμένου χαρακτήρα των οικονομικών και κοινωνικών συναλλαγών, πληθαίνουν οι ανάγκες ανταλλαγής δεδομένων προσωπικού χαρακτήρα μεταξύ δημόσιων και ιδιωτικών φορέων. Έτσι, κάθε κράτος μέλος της Ένωσης καλείται να ανταλλάσσει τέτοια δεδομένα προκειμένου να ανταπεξέλθει στις υποχρεώσεις του και στα καθήκοντά του, ακόμη και για λογαριασμό άλλου κράτους μέλους, στα πλαίσια παγιωμένης πλέον συνεργασίας.

Οι εξελίξεις στην τεχνολογία και η παγκοσμιοποίηση δημιούργησαν νέες προκλήσεις για την προστασία δεδομένων προσωπικού χαρακτήρα, καθώς η συλλογή και η ανταλλαγή δεδομένων έχουν κλιμακωθεί τόσο σε προσωπικό επίπεδο όσο και σε επίπεδο ιδιωτικών επιχειρήσεων και δημοσίων φορέων, με στόχο τη βελτιστοποίηση των δραστηριοτήτων τους.

Ακόμη, η νέα αυτή νομοθεσία θεσπίζει τους κανόνες βάσει των οποίων θα εφαρμόζεται έμπρακτα η προαναφερθείσα οδηγία, ενώ παράλληλα θα επιβληθούν νομικές και χρηματικές κυρώσεις για όσους παραβαίνουν τους κανόνες και δεν υιοθετούν την εν λόγω νομοθεσία.

Αναλυτικά, ο Γενικός Κανονισμός Προστασίας Δεδομένων (ΓΚΠΔ / GDPR) εκπόνηση σε πολυάριθμες νομικές διατάξεις και συνέταξε κανονισμούς που ενστερνίζονται οι ακόλουθες αρχές:

Η προστασία των φυσικών προσώπων στα πλαίσια της επεξεργασίας δεδομένων προσωπικού χαρακτήρα είναι, και πρέπει να θεωρείται, ένα από τα θεμελιώδη δικαιώματά τους.

Μέσω των διατάξεων του Κανονισμού, η Ευρωπαϊκή Ένωση στοχεύει στην ανάπτυξη ενός χώρου ελευθερίας, ασφάλειας και δικαιοσύνης, ανεξαρτήτως ιθαγενείας και τόπου διαμονής των ενδιαφερομένων φυσικών προσώπων.

Η οδηγία 95/46/ΕΚ, που εκπονήθηκε από το Ευρωπαϊκό Συμβούλιο, επιδίωκε την προστασία των θεμελιωδών δικαιωμάτων όσον αφορά τη διάδοση δεδομένων προσωπικού χαρακτήρα μεταξύ των κρατών μελών.

Η επεξεργασία δεδομένων και πληροφοριών τέτοιου τύπου οφείλει να εξυπηρετήσει ανθρώπινες ανάγκες. Η προστασία των ευαίσθητων δεδομένων δεν αποτελεί απόλυτο δικαίωμα αλλά χρειάζεται να εκτιμηθεί σε σχέση με τη λειτουργία του προσώπου στην κοινωνία.

Οι τελευταίες εξελίξεις στην τεχνολογία και η παγκοσμιοποίηση δημιούργησαν ισχυρές προκλήσεις για την προστασία των δεδομένων προσωπικού χαρακτήρα. Οι ανταλλαγές δεδομένων μεταξύ δημόσιων και ιδιωτικών φορέων έχουν εντατικοποιηθεί σημαντικά με σκοπό την προώθηση και τη βελτίωση των δραστηριοτήτων τους. Επιπλέον, τα ίδια τα φυσικά πρόσωπα δημοσιεύουν συνεχώς αυξανόμενο όγκο πληροφοριών ιδιωτικού και προσωπικού χαρακτήρα που, μέσω του διαδικτύου και των Social Media, είναι παγκόσμια προσβάσιμες.

Θεωρείται δυνατή και αναγκαία η ενίσχυση της οικονομίας χάρη στην πιστή εφαρμογή της νέας αυτής νομοθεσίας, η οποία προβλέπει επίσης τη δυνατότητα των κρατών μελών να την προσαρμόσουν στο εθνικό τους δίκαιο, ακόμη και με την εκπόνηση νέων νόμων με στόχο την εσωτερική ενσωμάτωση και εναρμόνιση της οδηγίας GDPR.

Η προηγούμενη οδηγία 95/46/ΕΚ δεν κατόρθωσε να αποτρέψει τον κατακερματισμό της εφαρμογής της προστασίας των δεδομένων σε ολόκληρη την Ένωση με αποτέλεσμα οι διαφορές όσον αφορά την επεξεργασία των δεδομένων να εμποδίσουν την άσκηση οικονομικών δραστηριοτήτων και να στρεβλώνουν τον ανταγωνισμό.

Για την αποτελεσματική προστασία των δεδομένων προσωπικού χαρακτήρα, χρειάζονται και απαιτούνται τόσο η ενίσχυση όσο και ο ενδελεχής καθορισμός των δικαιωμάτων των φυσικών προσώπων, καθώς και των υποχρεώσεών τους ως προς την επεξεργασία δεδομένων. Αναδύεται η ανάγκη ορισμού αντίστοιχων εξουσιών παρακολούθησης και διασφάλισης της συμμόρφωσης εκ μέρους εκάστοτε κράτους

μέλους προς τους κανόνες προστασίας αυτών των δεδομένων και επιβολής κυρώσεων για ενδεχόμενες παραβιάσεις. Ο κανονισμός δεν καλύπτει τα ζητήματα προστασίας δεδομένων που αφορούν νομικά πρόσωπα και, ειδικά, επιχειρήσεις συσταθείσες ως νομικά πρόσωπα.

Για να είναι τεχνολογικά άμεμπτη η προστασία των φυσικών προσώπων, θα πρέπει να εφαρμόζεται η οδηγία τόσο στην επεξεργασία δεδομένων με αυτοματοποιημένα μέσα, όσο και στη χειροκίνητη επεξεργασία, αν τα δεδομένα αυτά περιέχονται, ή προορίζονται να περιληφθούν, σε συστήματα αρχειοθέτησης.

Δεν απαιτείται η εφαρμογή του νόμου σε περιπτώσεις και δραστηριότητες που αφορούν ζητήματα εθνικής ασφάλειας, όπως επίσης και στις δραστηριότητες συναφείς με την κοινή εξωτερική πολιτική και την πολιτική ασφάλειας της Ένωσης.

Όταν η επεξεργασία αφορά αποκλειστικά το πλαίσιο προσωπικής ή οικείας δραστηριότητας φυσικού προσώπου, τότε δεν εφαρμόζεται ο κανονισμός αυτός, καθώς δε συνδέεται με κάποια επαγγελματική ή εμπορική δραστηριότητα. Τέτοια παραδείγματα είναι η αλληλογραφία και η κοινωνική δικτύωση. Ωστόσο, εφαρμόζεται για τους υπεύθυνους επεξεργασίας ή εκτελούντες την επεξεργασία οι οποίοι παρέχουν τα κατάλληλα τεχνικά μέσα.

Επίσης, όταν η επεξεργασία αφορά την πρόληψη, τη διερεύνηση, την ανίχνευση ή τη δίωξη ποινικών αδικημάτων και την επιβολή ποινικών κυρώσεων δεν εφαρμόζεται ο κανονισμός. Πάντως, στις εν λόγω περιπτώσεις τα κράτη μέλη πρέπει να υιοθετούν την ειδική νομική οδηγία (ΕΕ) 2016/680 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, ειδάλλως υπάγονται στο γενικό κανονισμό.

Για τη διευκόλυνση των υπευθύνων της επεξεργασίας προσωπικών δεδομένων και της έμπρακτης διενεργείας της, ο νόμος προβλέπει την υιοθέτηση της ψευδωνυμοποίησης[5], η οποία μπορεί να επιφέρει μείωση των κινδύνων δημοσιοποίησης των δεδομένων, ενώ ταυτόχρονα δεν αποκλείει την αξιοποίηση άλλων μέτρων προστασίας των δεδομένων. Για την υλοποίηση της τεχνικής αυτής, ο υπεύθυνος της επεξεργασίας προσωπικών δεδομένων, έχοντας αναλύσει τις συσχετιζόμενες πληροφορίες και ανάγοντας το συμπέρασμα ότι δεν υφίσταται πιθανότητα αναγνώρισης συγκεκριμένου φυσικού προσώπου, μπορεί να εξουσιοδοτήσει τους εκτελούντες τη διαδικασία να εφαρμόσουν το μέτρο της ψευδωνυμοποίησης. Με τον τρόπο αυτό δεν παραβιάζεται η οδηγία αυτή, ενώ παράλληλα επιτρέπεται η επεξεργασία των υπολοίπων πληροφοριών.

[5] Η τεχνική της ψευδωνυμοποίησης υιοθετείται από το ΓΚΠΔ και προτείνεται σε μερικές περιπτώσεις για την προστασία των ευαίσθητων πληροφοριών χωρίς να παραβιάζεται η νέα νομοθεσία.

Αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας

Η χρήση του διαδικτύου έχει επιφέρει νέες μεθόδους ανάπτυξης δραστηριοτήτων, όπως διαφημίσεις, που σχετίζονται με τις διαδικτυακές προτιμήσεις του ατόμου κ.α. Χαρακτηριστικό παράδειγμα αποτελούν τα cookies[6], τα οποία αποθηκεύουν-διατηρούν στοιχεία από τις επισκεψιμότητες των ιστοσελίδων εκ μέρους του χρήστη που, αν συνδυαστούν, δημιουργούν το λεγόμενο **"**προφίλ χρήστη**"** και μπορούν να οδηγήσουν σε ταυτοποίηση φυσικού προσώπου. Αυτός ο παράγοντας παραβιάζει το ΓΚΠΔ.

Το υποκείμενο οφείλει να διατυπώσει έμπρακτα την αποδοχή του ως προς την καθολική επεξεργασία των δεδομένων του, είτε με γραπτή ή προφορική δήλωση είτε επιλέγοντας ηλεκτρονικά το τετραγωνάκι που περιέχει τις τεχνικές ρυθμίσεις της αρεσκείας του ή, ακόμη, συντάσσοντας μια δήλωση από την οποία να προκύπτει σαφέστατα ότι συναινεί με την προτεινόμενη επεξεργασία δεδομένων προσωπικού χαρακτήρα. Ως εκ τούτου, τυχόν σιωπή του, με οποιοδήποτε τρόπο, εκλαμβάνεται ως άρνηση της πρότασης επεξεργασίας των ευαίσθητων δεδομένων του. Κατά τη διαβίβαση αιτήματος προς το υποκείμενο για συγκατάθεση, πρέπει να διευκρινισθούν λεπτομερώς οι σκοποί της επεξεργασίας. Όταν πρόκειται για επεξεργασία που εξυπηρετεί διαφόρους σκοπούς της επιστημονικής έρευνας, το υποκείμενο δύναται να συναινέσει για την επεξεργασία δεδομένων του μόνο για ορισμένους τομείς της.

Χαρακτηρίζονται δεδομένα προσωπικού χαρακτήρα και τα γενετικά δεδομένα, δηλαδή όσα κληρονομούνται και μεταβιβάζονται από γενιά σε γενιά και εντοπίζονται με εξετάσεις του DNA[7] ή του RNA[8]. Παράλληλα, συλλέγονται και προσωπικά δεδομένα που συγκαταλέγονται στο ιατρικό προφίλ του υποκείμενου, από τα οποία πρέπει να διαμορφωθεί πλήρης εικόνα της κατάστασης της υγείας του, τόσο της σωματικής όσο και της ψυχικής πτυχής. Η οδηγία 2011/24/ΕΕ ορίζει κάποιο διακριτικό στοιχείο που καθιστά εφικτή η απόλυτη ταυτοποίηση του φυσικού προσώπου για τη διευκόλυνση αναζητήσεων ιατρικής φύσης από οποιαδήποτε πηγή, σχετικά με νοσήματα, αναπηρίες, θεραπείες, ιατρικό ιστορικό και ούτως ως εξής.

Ξεχωριστή μέριμνα επιβάλλεται στην επεξεργασία προσωπικών δεδομένων ανηλίκων που, ως επί το πλείστων, αγνοούν κινδύνους, συνέπειες και δικαιώματα συνυφασμένα με την επεξεργασία προσωπικών δεδομένων, κυρίως όταν πρόκειται να χρησιμοποιηθούν για εμπορικούς σκοπούς. Δεν απαιτείται η συγκατάθεση του γονέα ή του κηδεμόνα στις περιπτώσεις παροχής υπηρεσιών ή συμβουλών στον ίδιο τον ανήλικο.

---

[6] Ένα «cookie» είναι μια μικρή ποσότητα δεδομένων που δημιουργούνται από έναν ιστότοπο και αποθηκεύονται από το πρόγραμμα περιήγησής σας. Σκοπός του είναι να θυμάται πληροφορίες για εσάς, παρόμοια με ένα αρχείο προτιμήσεων που δημιουργήθηκε από μια εφαρμογή λογισμικού. Ενώ τα «cookies» εξυπηρετούν πολλές λειτουργίες, ο συνηθέστερος σκοπός τους είναι η αποθήκευση πληροφοριών σύνδεσης για έναν συγκεκριμένο ιστότοπο.

[7] DNA είναι η αγγλική συντομογραφία που χρησιμοποιείται για την περιγραφή του Δεοξυριβονουκλεϊκού Οξέος.
[8] RNA είναι η αγγλική συντομογραφία που χρησιμοποιείται για την περιγραφή του Ριβονουκλεϊκού Οξέος.

Η νέα υπό εξέταση οδηγία εφιστά την προσοχή στην ενδεχόμενη μεταβολή των σκοπών για τους οποίους πραγματοποιείται η επεξεργασία προσωπικών δεδομένων και ορίζει ότι είναι νόμιμη μόνο με την προϋπόθεση σύμπνοιας των αρχικών σκοπών και των μεταγενέστερων σκοπών, οπότε στηρίζονται στην ίδια νομική βάση. Έννομη θεωρείται η επεξεργασία όταν επιβάλλεται από λόγους δημοσίου συμφέροντος και διενεργείται από ειδικά εξουσιοδοτημένο άτομο, ενώ η κοινοτική ή η εθνική νομοθεσία καθορίζουν συνάμα τους σκοπούς της επεξεργασίας και τα καθήκοντα του υπευθύνου, όπως και στην περίπτωση περαιτέρω επεξεργασίας δεδομένων προς το δημόσιο συμφέρον, για επιστημονική ή ιστορική έρευνα ή για στατιστικούς σκοπούς. Ο υπεύθυνος επεξεργασίας οφείλει να σταθμίσει διαφόρους παράγοντες όπως, παραδείγματος χάρη, τη συνάφεια των αρχικών σκοπών με εκείνους της περαιτέρω επεξεργασίας, το πλαίσιο συγκέντρωσης των ευαίσθητων δεδομένων, τις προσδοκίες του υποκειμένου των δεδομένων κ.α.

Άλλη σημαντική καινοτομία του παρόντος κανονισμού έγκειται στη δυνατότητα που παρέχεται στον υπεύθυνο επεξεργασίας να προχωρήσει νόμιμα, μετά από τη συγκατάθεση του κατόχου των δεδομένων, σε περαιτέρω επεξεργασία των ευαίσθητων δεδομένων, όταν διακυβεύονται δημόσια συμφέροντα· αδιαμφισβήτητος κανόνας είναι η τήρηση της αρχής της ενημέρωσης του κατόχου των δεδομένων σχετικά με τους άλλους σκοπούς της περαιτέρω επεξεργασίας, σχετικά επίσης με τα δικαιώματά του, χωρίς να παραγκωνιστεί το δικαίωμα εκδήλωσης αντιρρήσεων.

Θεμιτό συμφέρον θεωρείται επίσης η διαβίβαση εκ μέρους του υπεύθυνου επεξεργασίας ευαίσθητων δεδομένων που συσχετίζονται με εγκληματικές πράξεις ή απειλές για τη δημόσια ασφάλεια. Ωστόσο, η εν λόγω διαβίβαση απαγορεύεται όταν αντιβαίνει σε θεσπιζόμενη από διατάξεις του νόμου υποχρέωση τήρησης εχεμύθειας.

Ακόμη και στις περιπτώσεις αρχειοθέτησης που εξυπηρετεί δημόσιο συμφέρον ή όταν συντρέχουν λόγοι συνυφασμένοι με την επιστημονική, ιστορική ή στατιστική έρευνα, κατά την επεξεργασία προσωπικών δεδομένων θα πρέπει να παρασχεθούν εγγυήσεις απόλυτης συμμόρφωσης με τις αρχές αυτού του κανονισμού και υιοθέτησης των πλέον ενδεδειγμένων τεχνικών μέσων. Στην περίπτωση που, για τους ίδιους ως άνω σκοπούς, ο υπεύθυνος της επεξεργασίας κρίνει σκόπιμο να προχωρήσει σε πλέον εμπεριστατωμένη επεξεργασία, απαραίτητες προϋποθέσεις είναι η αδυναμία ταυτοποίησης των κατόχων των δεδομένων και η παροχή εγγυήσεων, μεταξύ των άλλων, με την ψευδωνυμοποίηση. Δεν πρέπει να παραβλεφθεί ο ρόλος των κρατών μελών στη διασφάλιση των δικαιωμάτων στα πλαίσια της επεξεργασίας προσωπικών δεδομένων, στον καθορισμό προδιαγραφών και αποκλίσεων, στη διασφάλιση του δικαιώματος περιορισμού της επεξεργασίας, μέσα από τη διάρθρωση στοχευμένων διαδικασιών. Ιδιαίτερη μνεία χρήζουν οι κλινικές δοκιμές για τις οποίες η επεξεργασία δεδομένων θα πρέπει να υιοθετήσει συγκεκριμένες νομοθετικές διατάξεις.

Εξετάζοντας την εφαρμογή του νέου κανονισμού ως προς τον τόπο διενέργειας της επεξεργασίας προσωπικών δεδομένων, επισημαίνεται ότι ο παράγοντας αυτός δεν

εντάσσεται σε αυστηρά καθορισμένα πλαίσια. Ο υπεύθυνος της επεξεργασίας μπορεί να είναι εγκατεστημένος και εκτός Ευρωπαϊκής Ένωσης ακόμη και όταν τα δεδομένα αφορούν συναλλαγές με υποκείμενα δεδομένων ή συμπεριφορές τους εντός της Ένωσης και στην περίπτωση αυτή υιοθετείται το δίκαιο κράτους μέλους, όπως ορίζει το διεθνές δίκαιο. Εξέχουσα σημασία έχουν, αντίθετα, οι επεξηγήσεις των όρων που καθιερώνονται στο νέο κανονισμό, όπως για παράδειγμα: **"**δεδομένα προσωπικού χαρακτήρα**",** **"**επεξεργασία**",** **"**περιορισμός επεξεργασίας**",** **"**κατάρτιση προφίλ**",** **"**ψευδωνυμοποίηση**",** **"**σύστημα αρχειοθέτησης**",** **"**υπεύθυνος επεξεργασίας**",** **"**εκτελών την επεξεργασία**",** **"**αποδέκτης**",** **"**τρίτος**",** **"**συγκατάθεση**",** **"**παραβίαση δεδομένων προσωπικού χαρακτήρα**",** **"**γενετικά δεδομένα**",** **"**βιομετρικά δεδομένα**",** **"**δεδομένα που αφορούν την υγεία**"** κ.α. Κάθε μεμονωμένος όρος, που ο νέος κανονισμός αναλύει σε βάθος, αποτελεί αναντικατάστατο εργαλείο για την επίτευξη του στόχου της προστασίας των ευαίσθητων δεδομένων, στη διάθεση τόσο του υποκειμένου όσο και του υπεύθυνου της επεξεργασίας.

Δεν χωρεί ουδεμία αμφιβολία ότι η προστασία των προσωπικών δεδομένων ανήκει στη σφαίρα των ανθρωπίνων δικαιωμάτων και για το λόγο αυτό ο νέος κανονισμός εστιάζει την προσοχή του στις αρχές που θεμελιώνουν την επεξεργασία των προσωπικών δεδομένων. Οι βασικές αυτές αρχές είναι συνοπτικά: η νομιμότητα, η αμεροληψία, η διαφάνεια, η ελαχιστοποίηση, η στοχοποίηση, η ακρίβεια.

Καθορίζεται ακόμη και η χρονική περίοδος διατήρησης των δεδομένων, περιορίζοντάς την στον απολύτως απαραίτητο χρόνο για την επίτευξη των στόχων της επεξεργασίας. Τυχόν χρονική παράταση προβλέπεται σε περίπτωση αρχειοθέτησης που εξυπηρετεί δημόσιο συμφέρον και επιστημονικούς σκοπούς.

Με τον όρο **"**ακεραιότητα και εμπιστευτικότητα**"** διασφαλίζεται η εχεμύθεια των δεδομένων προσωπικού χαρακτήρα, η προστασία τους από παράνομες παρεμβάσεις και η υιοθέτηση προληπτικών μέτρων για την αποφυγή οποιασδήποτε μορφής ζημιών των δεδομένων.

Με άλλες διατάξεις του νέου κανονισμού θεσπίζονται ζητήματα που αφορούν πτυχές, όπως η νομιμότητα της επεξεργασίας των δεδομένων, άρρηκτα συνδεδεμένη με τη συναίνεση του υποκειμένου των δεδομένων, η αναγκαιότητα της επεξεργασίας δεδομένων όταν πρόκειται για συμβαλλόμενα μέρη, ακόμη και πριν από τη σύναψη συμβάσεων, η πιστή τήρηση των νόμιμων υποχρεώσεων του υπευθύνου επεξεργασίας, η διασφάλιση των συμφερόντων του υποκειμένου των δεδομένων και της δημόσιας διοίκησης και η προστασία των ανθρωπίνων δικαιωμάτων, ιδίως των ανηλίκων.

Σε κάθε περίπτωση, παρέχεται στα κράτη μέλη της Ευρωπαϊκής Ένωσης το αδιαφιλονίκητο δικαίωμά τους να εισαγάγουν στη νέα νομοθεσία για την προστασία των προσωπικών δεδομένων εκείνες τις διατάξεις που κρίνονται απαραίτητες καθώς καθρεφτίζουν την πραγματικότητα και την ιδιαιτερότητά τους, συμπληρώνοντας με τον τρόπο αυτό το γενικό δίκαιο της ΕΕ.

Με το ίδιο σκεφτικό χαράζονται οι σκοποί της προστασίας των δεδομένων προσωπικού χαρακτήρα και του ιδίου του υποκειμένου των δεδομένων αυτών. Σε περίπτωση απόκλισης των σκοπών της επεξεργασίας από το δίκαιο της Ευρωπαϊκής Ένωσης ή των μεμονωμένων κρατών μελών ή έλλειψης συγκατάθεσης του υποκειμένου, η συμβατότητά τους τεκμηριώνεται αναλόγως της σχέσης των σκοπών για την αρχική επεξεργασία και εκείνης της περαιτέρω επεξεργασίας και της σχέσης μεταξύ υποκειμένων και υπευθύνου επεξεργασίας. Εξετάζονται επίσης το είδος των ευαίσθητων δεδομένων, τυχόν συνέπειες περαιτέρω επεξεργασίας και η παροχή εγγυήσεων.

Άλλο πρωταρχικής σημασίας ζήτημα είναι εκείνο της συγκατάθεσης για την επεξεργασία δεδομένων προσωπικού χαρακτήρα, η οποία πρέπει να αποδειχθεί και το σχετικό υποβληθέν αίτημα εκ μέρους του υπευθύνου επεξεργασίας πρέπει να ξεχωρίζει σαφώς και ευκολονόητα από άλλα θέματα που περιέχονται στην ίδια δήλωση. Η συγκατάθεση μπορεί να ανακληθεί από το υποκείμενο των δεδομένων, ενώ η επεξεργασία που έχει ήδη διενεργηθεί παραμένει έγκυρη. Η παροχή συγκατάθεσης πρέπει να είναι απρόσκοπτη. Όσον αφορά τη συγκατάθεση επεξεργασίας δεδομένων που αφορούν ανηλίκου και προς όφελος του ιδίου, απαραίτητη προϋπόθεση είναι η συμπλήρωση του 16^{ου} έτους ηλικίας ή, σε εξαιρετικές περιπτώσεις, το του 13^{ου} έτους ηλικίας. Στην αντίθετη περίπτωση αρμόδιο για την παροχή της συγκατάθεσης είναι το άτομο που ασκεί τη γονική μέριμνα, ενώ ο υπεύθυνος επεξεργασίας οφείλει να εξακριβώσει τη νομιμότητα της συγκατάθεσης.

Ο νέος κανονισμός απαγορεύει την επεξεργασία προσωπικών δεδομένων όταν αναδύονται απ' αυτά φυλετικά στοιχεία ή ιδεολογικές και πολιτικές πεποιθήσεις, όπως επίσης γενετικά στοιχεία, ιατρικές και σεξουαλικές πληροφορίες. Η απαγόρευση αυτή δεν ισχύει όταν το υποκείμενο συναινεί απόλυτα με την επεξεργασία, όταν διακυβεύονται δικαιώματα του υπευθύνου επεξεργασίας και του υποκειμένου των δεδομένων και υπαγορεύεται από λόγους προστασίας τω συμφερόντων ατόμων με σωματικές ή νομικές ανικανότητες. Η απαγόρευση παρακάμπτεται επίσης όταν την ευθύνη της επεξεργασίας αναλαμβάνει κάποιος επαγγελματίας που δεσμεύεται από το επαγγελματικό απόρρητο.

Ο νέος κανονισμός προστατεύει τα δικαιώματα υποκειμένων των δεδομένων προσωπικού χαρακτήρα, ακόμη και όταν έχουν διαπράξει αδικήματα και έχουν εκδοθεί σε βάρος τους ποινικές καταδίκες, υποχρεώνοντας τις αρμόδιες Αρχές να προβούν στον έλεγχο της επεξεργασίας των προσωπικών δεδομένων τους.

Ένα άλλο άρθρο του νέου κανονισμού ρυθμίζει το ζήτημα της εξακρίβωσης της ταυτότητας του υποκειμένου των δεδομένων στα πλαίσια της επεξεργασίας τους. Συγκεκριμένα, προβλέπει ότι η υποχρέωση του υπευθύνου της επεξεργασίας να προβεί στην εν λόγω εξακρίβωση εξαρτάται από τους σκοπούς της ίδιας της επεξεργασίας, ενημερώνοντας σχετικά το υποκείμενο των δεδομένων. Αν η εξακρίβωση της

ταυτότητας είναι αδύνατη λόγω ελλιπών στοιχείων ο υπεύθυνος επεξεργασίας μπορεί να ζητήσει συμπληρωματικές πληροφορίες.

Ο υπεύθυνος επεξεργασίας οφείλει να ενημερώσει δωρεάν, τακτικά, με σαφήνεια και διαφάνεια, το υποκείμενο των δεδομένων για όλες τις ενέργειες και τις εξελίξεις της επεξεργασίας των δεδομένων του. Ο υπεύθυνος επεξεργασίας μπορεί να επιβάλει κάποια επιβάρυνση όταν τα αιτήματα του υποκειμένου είναι αποδεδειγμένα αστήριχτα ή υπέρμετρα. Μπορεί επίσης να αρνηθεί να τα ικανοποιήσει. Όταν είναι το ίδιο το υποκείμενο των δεδομένων που συγκεντρώνει τα προσωπικά στοιχεία του στον υπεύθυνο επεξεργασίας, ο τελευταίος του διαβιβάζει πλήρη στοιχεία επικοινωνίας είτε μαζί του είτε με κάποιο εκπρόσωπό του, ενημερώνοντας τον για τους σκοπούς της επεξεργασίας και το νομικό πλαίσιο της, όπως και για τα στοιχεία επικοινωνίας με τον υπεύθυνο προστασίας των δεδομένων. Το υποκείμενο δεδομένων δικαιούται να ενημερώνεται, με σκοπό να εγγυηθεί η νομιμότητα και η διαφάνεια της επεξεργασίας, για ενδεχόμενη κοινοποίησή τους σε άλλο κράτος ή σε διεθνή οργανισμό, για την χρονική διάρκεια αποθήκευσης των δεδομένων του καθώς και τα κριτήρια αποθήκευσης, για τυχόν δυνατότητά του να επιφέρει τροποποιήσεις παντός τύπου στα δεδομένα του και, τέλος, σχετικά με τη φορητότητά τους. Ένα άλλο δικαίωμα που παγιοποιήθηκε στο νέο υπό εξέταση κανονισμό αφορά τη δυνατότητα του υποκείμενου των δεδομένων να ανακαλέσει ανά πάσα στιγμή τη συγκατάθεσή του για επεξεργασία, με παράλληλη δυνατότητα καταγγελίας σε αρμόδιο για την εποπτεία της επεξεργασίας Όργανο.

Διαχωρίζονται στη νέα οδηγία οι νομικές από τις συμβατικές υποχρεώσεις διαβίβασης ευαίσθητων δεδομένων με την πρόβλεψη των συνεπειών συνυφασμένων με την άρνηση παροχής τους.

Στην αντίθετη περίπτωση, όταν δηλαδή τα δεδομένα προσωπικού χαρακτήρα δεν συγκεντρώνονται από το ίδιο το υποκείμενο, ο υπεύθυνος επεξεργασίας διαβιβάζει στο υποκείμενο τα στοιχεία για την ταυτοποίηση του και για την μεταξύ τους επικοινωνία ή για επικοινωνία με τον εκπρόσωπό του. Για τις λοιπές ρυθμίσεις ισχύουν τα όσα προβλέπονται για τις προαναφερόμενες περιπτώσεις.

Ο υπεύθυνος επεξεργασίας διαβιβάζει τις παραπάνω πληροφορίες το αργότερο εντός ενός μηνός από τη συλλογή των δεδομένων ή με την πρώτη επικοινωνία με το υποκείμενο των δεδομένων ή με ενδεχόμενο τρίτο αποδέκτη.

Όταν πρόκειται για περαιτέρω επεξεργασία το υποκείμενο δεδομένων οφείλει να γνωρίζει αν υφίσταται και ποιος είναι ο σκοπός της πριν από τη διενέργειά της. Ο υπεύθυνος επεξεργασίας, στα πλαίσια της συγκεκριμένης περαιτέρω επεξεργασίας, οφείλει να μεριμνήσει για την προστασία των δικαιωμάτων του υποκειμένου των δεδομένων, ιδίως όταν η συγκέντρωση πληροφοριών παρουσιάζεται λίαν δυσχερής ή ακόμη και επιζήμια για το δημόσιο συμφέρον και για τους σκοπούς της αρχειοθέτησης. Στον ίδιο κοινό παρονομαστή εντάσσονται η συγκέντρωση και η δημοσιοποίηση δεδομένων που προβλέπονται στο κοινοτικό δίκαιο ή στην εθνική νομοθεσία κράτους

μέλους, όπως καις την περίπτωση διαφύλαξης του απορρήτου δεδομένων επαγγελματικής φύσης.

Στο ευρύτατο φάσμα των ζητημάτων που αναλύονται και λύνονται με τις ρυθμίσεις του νέου κανονισμού προβάλλεται και εκείνο που αφορά το δικαίωμα πρόσβασης εκ μέρους του υποκειμένου των δεδομένων, για το οποίο θεσπίζεται το πάγιο δικαίωμα να γνωρίζει αν τα προσωπικά του δεδομένα αποτελούν αντικείμενο επεξεργασίας και να έχει πρόσβαση στα εν λόγω δεδομένα, ώστε να διαπιστώσει τους σκοπούς της επεξεργασίας, ποιοι είναι οι αποδέκτες των δεδομένων, ειδικά όταν δρουν στο εξωτερικό ή είναι οργανισμοί διεθνούς εμβέλειας. Το υποκείμενο των δεδομένων δικαιούται επίσης να γνωρίζει το χρονικό διάστημα διάρκειας της αποθήκευσης των δεδομένων και να απαιτήσεις σχετικές εγγυήσεις, να υποβάλει αίτημα για διόρθωση ή τροποποίηση των δεδομένων, να προβεί σε τυχόν καταγγελία προς τον αρμόδιο για την εποπτεία φορέα. Το υποκείμενο των δεδομένων δικαιούται να λάβει αντίγραφο των επεξεργαζόμενων δεδομένων από τον υπεύθυνο επεξεργασίας. Για παροχή περισσότερων αντιγράφων ο υπεύθυνος επεξεργασίας μπορεί να επιβάλει την πληρωμή τέλους για διοικητικές δαπάνες, όπως προαναφέρθηκε.

Το υποκείμενο των δεδομένων προσωπικού χαρακτήρα δικαιούται επίσης να ζητήσει από τον υπεύθυνο επεξεργασίας να διορθώσει ή να συμπληρώσει τυχόν λάθη ή ελλείψεις των δεδομένων, όπως και τη διαγραφή τους όταν δεν παύει πλέον η αναγκαιότητά τους και δεν εξυπηρετούν πλέον τους σκοπούς για τους οποίους είχαν συγκεντρωθεί ή όταν ανακαλεί τη συγκατάθεσή του για επεξεργασία. Αν η επεξεργασία δεν στηρίζεται σε επιταγές του νόμου, το υποκείμενο των δεδομένων μπορεί να αντιταχθεί στην ίδια την επεξεργασία.

Η διαγραφή των δεδομένων επιβάλλεται για την τήρηση νομικών υποχρεώσεων, όταν η συγκέντρωσή τους ήταν άμεσα συνδεδεμένη με την προσφορά υπηρεσιών πληροφόρησης και όταν το ζητήσει ρητά το υποκείμενο των δεδομένων λόγω ύπαρξης συνδέσμων ή αντιγράφων ή αναπαραγωγής τους. Τα όσα προαναφέρθηκαν περί διαγραφής των δεδομένων προσωπικού χαρακτήρα δεν ισχύουν όταν προβάλλονται δικαιώματα ελεύθερης έκφρασης και ενημέρωσης, όταν υφίστανται λόγοι τήρησης νομικών υποχρεώσεων, στα πλαίσια του δικαίου της Ένωσης ή κάποιου κράτους μέλους, για την προστασία δημοσίου συμφέροντος και για σκοπούς αρχειοθέτησης, όπως και για νομικές διεκδικήσεις.

Ένα σημαντικό μέτρο που θεσπίζει η οδηγία εκ μέρους των φυσικών ατόμων, αφορά τη δυνατότητα κάθε υποκειμένου να περιορίσει την επεξεργασία δεδομένων που το αφορούν σε περιπτώσεις που αμφισβητείται η ακρίβειά τους και όταν η χρήση των δεδομένων αυτών κρίνεται παράνομη. Εάν το φυσικό άτομο διαφωνεί με την επεξεργασία, μπορεί να καταθέσει ένσταση, ώστε να επαληθευτεί ή όχι σε αντίθετη περίπτωση ο ισχυρισμός του. Στο διάστημα αυτό υφίσταται προσωρινός περιορισμός των εν λόγω δεδομένων και πληροφοριών. Επίσης, η οδηγία ορίζει την υποχρέωση γνωστοποίησης από πλευράς των υπεύθυνων επεξεργασίας, σε περιπτώσεις αλλαγών,

διορθώσεων επεξεργασίας, διαγραφών ή περιορισμών επεξεργασίας, σε όλους τους αποδέκτες που γνωστοποιήθηκαν οι πληροφορίες αυτές. Επιπρόσθετα, ένα καθήκον του υπεύθυνου επεξεργασίας, είναι η έγκυρη και η άμεση πληροφόρηση του φυσικού ατόμου για τους παραπάνω αποδέκτες των πληροφοριών, εάν ζητηθεί από το εκάστοτε υποκείμενο.

Τα φυσικά άτομα, έχουν κάθε δικαίωμα στη φορητότητα των δεδομένων αυτών, δηλαδή, αρχικά να τα λαμβάνει από τον υπεύθυνο επεξεργασίας χωρίς καμία αντίσταση από τον τελευταίο και στη συνέχεια αν το επιθυμεί τη διαβίβασή τους σε κάποιον άλλο υπεύθυνο επεξεργασίας. Αυτό συμβαίνει όταν έχει προηγηθεί συγκατάθεση ή σύμβαση και ταυτόχρονα η επεξεργασία διενεργείται με αυτοματοποιημένα μέσα. Το φυσικό άτομο σε όλες τις παραπάνω περιπτώσεις, έχει δικαίωμα να αντιτάσσεται, οποτεδήποτε κρίνει ο ίδιος αναγκαίο, στην επεξεργασία των δεδομένων που το αφορούν. Εάν οι πληροφορίες που υποβάλλονται σε επεξεργασία έχουν στόχο την εμπορική προώθηση, τότε το υποκείμενο των πληροφοριών αυτών, έχει κάθε δικαίωμα να αντιταχθεί και τα δεδομένα αυτά να μην υποβάλλονται πλέον σε επεξεργασία για τους παραπάνω στόχους.

Σε ειδικές περιπτώσεις, όπου η επεξεργασία έχει στόχους επιστημονικών και ιστορικών ερευνών ή στατιστικούς σκοπούς, τότε υποκείμενο έχει επίσης δικαίωμα να αντιταχθεί για προσωπικούς του λόγους επεξεργασία των δεδομένων που το αφορούν εκτός αν αυτή κρίνεται απαραίτητη για την εκτέλεση καθηκόντων με δημόσιο συμφέρον.

2.2 Προσέγγιση Εργασίας και Τεχνικές προσαρμογής στην οδηγία

Οι παραπάνω οδηγίες που ορίζει ο Γενικός Κανονισμός Προστασίας Δεδομένων, αναγκάζει τους φορείς δημόσιους και ιδιωτικούς να προσαρμοστούν άμεσα στις νέες αναγκαίες μεταρρυθμίσεις που απαιτούνται για να συνεχίσουν να αναπτύσσουν τις δραστηριότητές τους χωρίς χρηματικές και ποινικές κυρώσεις. Ειδικότερα, οι φορείς που ασχολούνται άμεσα και έμμεσα με τον κλάδο της Πληροφορικής συναντούν πολλές δυσκολίες προσαρμογής, αφού ο κανονισμός είναι αρκετά απαιτητικός και σε αρκετές περιπτώσεις δυσνόητος για τους περισσότερους που αναπτύσσουν τις δραστηριότητές τους μέσω των δεδομένων.

Έχοντας υπόψη τα όσα προαναφέρθηκαν, ο Γενικός Κανονισμός Προστασίας Δεδομένων επηρεάζει τις υφιστάμενες διαδικασίες αρχειοθέτησης σε όλους τους φορείς που δρουν στην Ευρωπαϊκή Κοινότητα. Συγκεκριμένα, με την εφαρμογή του ΓΚΠΔ, όλες οι ενέργειες αποθήκευσης και επεξεργασίας δεδομένων προσωπικού χαρακτήρα ελέγχονται λεπτομερώς από τον Υπεύθυνο Προστασίας Δεδομένων (Data Protection Officer[9]) εκάστοτε φορέα. Η νέα αυτή θέση εργασίας έχει δημιουργηθεί ως αναγκαίο αποτέλεσμα της νομοθεσίας, αφού κρίνεται απαραίτητη η εμπεριστατωμένη μελέτη και κατανόηση των κανονισμών της και, παράλληλα, η υλοποίηση πρακτικών

διαδικασιών και μεθόδων για την αποφυγή της παραβίασης της οδηγίας αυτής που επιφέρει χρηματικές ποινές και ποινικές διώξεις.

Μέσα από πολλά άρθρα, διαπιστώνεται ότι, κάποιος σε αυτή τη νέα θέση εργασίας, πρέπει να έχει πλήρη γνώση και κατανόηση του νέου νόμου και παράλληλα να έχει τις απαραίτητες γνώσεις για τα εκάστοτε δεδομένα που εισέρχονται στον φορέα, τους στόχους και τους σκοπούς μιας επιχείρησης και αν χρησιμοποιεί τις νέες τεχνολογίες για την επίτευξη νέων στόχων, όπως είναι η Τεχνητή Νοημοσύνη, η Εξόρυξη Δεδομένων, η Μηχανική Μάθηση, που έχουν ενταχθεί στην καθημερινότητά μας και είναι κύριος στόχος κάθε φορέα να τις εντάξει στις δραστηριότητές του, για την επίτευξη στόχων, την βελτιστοποίηση των καθηκόντων, των αρμοδιοτήτων, την πολιτική της.

Είναι πλέον γνωστό, ότι κάθε δεδομένο «κρύβει» και μια πληροφορία, πόσο μάλλον πολλά δεδομένα οδηγούν σε πληροφορίες που μπορεί να ήταν άγνωστες, ή σε νέες πληροφορίες και συμπεράσματα που «το μάτι» δεν μπορεί να διακρίνει. Το θέμα που είχε δημιουργηθεί πριν την οδηγία, είναι η εκμετάλλευση κάθε είδους δεδομένων και γενικότερα πληροφοριών απαξιώντας εντελώς αν γίνεται κατάχρηση της πληροφορίας πόσο μάλλον όταν πρόκειται για ευαίσθητες πληροφορίες.

Με τον όρο ευαίσθητη πληροφορία, νοείται οποιοδήποτε δεδομένο μπορεί να οδηγήσει στην ταυτοποίηση φυσικού ατόμου άμεσα ή έμμεσα. Η οδηγία αυτή, έχει στόχο την τήρηση της προσωπικής ελευθερίας και την ασφάλεια των ευαίσθητων προσωπικών δεδομένων, ενώ με σαφήνεια, έχοντας γνώση τα οφέλη των νέων τεχνολογιών, συνιστά την συνέχιση των επιμέρους δραστηριοτήτων με προσαρμογή, με επίγνωση και με συνείδηση.

Για να γίνει πιο κατανοητό στην πράξη, ακολουθεί ένα καθημερινό παράδειγμα. Μια πληροφορία ή κάποιο δεδομένο μπορεί να είναι εν μέρει ευαίσθητο ή προσωπικό, οπότε και προβλέπεται η ασφάλεια μόνο του συγκεκριμένου μέρους της πληροφορίας και όχι ολόκληρη η πληροφορία. Για παράδειγμα, ένας λογαριασμός ηλεκτρονικού ταχυδρομείου πχ aimiliosvatikiotis@gmail.com, μπορεί στο πρώτο μέρος του email, να εμφανίζεται το όνομα ή το επίθετο ή ένας συνδυασμός αυτών που υπόκειται σε φυσικό άτομο όπως εδώ το «aimiliosvatikiotis», ενώ το δεύτερο μέρος «@gmail.com», να μη θεωρείται ευαίσθητη πληροφορία.

---

[9] Αποτέλεσμα του GDPR είναι η θέσπιση και η ίδρυση μιας νέας θέσης εργασίας στην οικονομία που έχει ως τίτλο «DPO» (Data Protection Officer), όπου θεωρείται απαραίτητη η πλήρης γνώση της νομοθεσίας και ως καθήκοντα αυτής της θέσης είναι η παρακολούθηση, η παρατήρηση, η επίβλεψη και η υποβολή του κανονισμού όπου θεωρείται αναγκαίο, σε φορείς και οργανισμούς με συστήματα αρχειοθέτησης για την αποφυγή χρηματικών και ποινικών κυρώσεων.

Σύμφωνα και με την οδηγία, μια απλή εναλλαγή χαρακτήρων στα προσωπικά δεδομένα, πολλές φορές δεν είναι επαρκής και επιβάλλεται ο μετασχηματισμός της πληροφορίας με τις εξής μεθόδους:

α) Ψευδωνυμοποίηση (Pseudo – Anonymization)

Διευκολύνει την επεξεργασία δεδομένων πέρα από τους αρχικούς σκοπούς αρχειοθέτησής τους (συνήθως διατήρηση).
Για να επιτευχθεί αυτό, υπάρχουν τρείς τεχνικές:

1. μετασχηματίζουμε κατάλληλα τις πληροφορίες αυτές σε τυχαίες συμβολοσειρές
2. Κρυπτογράφηση με SHA 256/512 bits[10]
3. Ψηφιακό κλειδί[11]

β) Ανωνυμοποίηση (Anonymization)

Οδηγεί σε μη αναστρέψιμη απομάκρυνση πληροφοριών που θα μπορούσαν να καταλήξουν σε συγκεκριμένα άτομα, είτε από αφαιρεμένες πληροφορίες, είτε σε συνδυασμό με άλλες πληροφορίες. Τα ανώνυμα δεδομένα πρέπει να αφαιρούνται από αναγνωρίσιμες πληροφορίες με σκοπό την ασφάλεια των ατόμων. Όταν γίνει σωστά, τότε η ανωνυμοποίηση τοποθετεί την επεξεργασία και την αποθήκευση δεδομένων εκτός του πεδίου εφαρμογής του GDPR.

γ) Καταστολή (Suppression)

Η καταστολή ή η αποκρυπτογράφηση δεδομένων είναι μια ακραία μορφή ανωνυμοποίησης.

Πως εφαρμόζεται αυτή η τεχνική:

Με τη μετατροπή των πληροφοριών σε κείμενο.

Η αποκρυπτογράφηση δεδομένων είναι πολύ απλή στην εφαρμογή και είναι αποτελεσματική στην απομάκρυνση των ευαίσθητων δεδομένων. Απ’ την άλλη πλευρά, τυχόν στατιστική ή αναλυτική αξία δεδομένων χάνεται στην διαδικασία.

---

[10] SHA είναι η αγγλική συντομογραφία του όρου Secure Hashing Algorithms, ενώ τα 256 ή 512 bits ισοδυναμούν με το μέγεθος του αποτελέσματος που θα έχει το αποτέλεσμα της κρυπτογράφησης της αρχικής πληροφορίας, δηλαδή το κρυπτογραφημένο τελικό αρχείο. Είναι μια από τις ασφαλέστερους αλγορίθμους κατακερματισμού στην αγορά.

[11] Το ψηφιακό κλειδί χρησιμοποιείται συνήθως για την έκδοση και διαχείριση ψηφιακών πιστοποιητικών (digital certificates) και από μεθόδους κρυπτογράφησης πληροφορίας ώστε να εξασφαλίζεται η ασφαλή λειτουργία των εξειδικευμένων προϊόντων λογισμικού και υλικού από αρμόδιους και εξουσιοδοτημένους χρήστες. Ως παράδειγμα μπορούμε να πούμε ότι τα δημόσια κλειδιά χρησιμοποιούνται από εξυπηρετητές και τα ιδιωτικά κλειδιά αντίστοιχα από εξουσιοδοτημένους χρήστες.

δ) Γενίκευση (Generalization)

Σε αυτή τη μέθοδο, οι ατομικές τιμές των πεδίων αποκαθίστανται με μία ευρύτερη κατηγορία, πχ: σε διαστήματα τιμών
Αυτή η κατηγορία δεν υπόκεινται σε περιορισμούς GDPR.

Είναι μια μη αναστρέψιμη διαδικασία και τα δεδομένα δε μπορούν να αποκατασταθούν στην αρχική μορφή που μπορεί να έχουν αναγνωριστεί.

Αυτή η μέθοδος αν συνδυαστεί με την καταστολή (suppression) τότε οδηγούμαστε σε μία νέα μέθοδο που αποκαλείται k – anonymity.

Η k – anonymity, είναι ουσιαστικά μια απόπειρα επαναχρησιμοποίησης δεδομένων που αφορούν συγκεκριμένα άτομα, παράγοντας δεδομένα με επιστημονική εγγύηση, ότι τα άτομα που αποτελούν τα δεδομένα δε γίνεται να επαναπροσδιοριστούν ενώ παράλληλα τα δεδομένα παραμένουν πρακτικά χρήσιμα.

ε) Κρυπτογράφηση (Encryption)

Η κρυπτογράφηση μεταφράζει τα δεδομένα σε μια άλλη μορφή, ώστε να μπορούν να την διαβάσουν μόνο οι άνθρωποι ή ένα σύστημα με πρόσβαση σε ένα μυστικό κλειδί (που ονομάζεται επίσημα "κλειδί αποκρυπτογράφησης"). Σύμφωνα με το άρθρο 32 του GDPR, οι υπεύθυνοι επεξεργασίας υποχρεούνται να εφαρμόζουν μέτρα βάσει κινδύνου για την προστασία της ασφάλειας των δεδομένων. Ένα τέτοιο μέτρο είναι η "κρυπτογράφηση δεδομένων προσωπικού χαρακτήρα" που "καθιστά τα δεδομένα ακατανόητα σε οποιοδήποτε πρόσωπο που δεν έχει εξουσιοδότηση πρόσβασης σε αυτό".

Οι οργανισμοί μπορούν να χρησιμοποιούν κρυπτογράφηση για να ικανοποιήσουν τις απαιτήσεις ασφάλειας δεδομένων GDPR. Υπάρχουν δύο σημαντικά προγράμματα κρυπτογράφησης δεδομένων: το συμμετρικό σύστημα κρυπτογράφησης και το ασύμμετρο σύστημα κρυπτογράφησης. Συχνά, αυτά τα δύο συστήματα αναμειγνύονται μαζί, και στη συνέχεια ονομάζεται "υβριδικό σύστημα κρυπτογράφησης".

Το σημερινό πρότυπο συμμετρικού συστήματος κρυπτογράφησης είναι το AES[12]. Είναι πολύ γρήγορο, συχνά επιταχύνεται από έναν επεξεργαστή (CPU). Για το λόγο αυτό έχει αμελητέο αντίκτυπο στην απόδοση του συστήματος. Η συμμετρική κρυπτογράφηση έχει την αδυναμία της παρουσία του μυστικού κλειδιού στην πλευρά κρυπτογράφησης, η οποία είναι εγγενώς δύσκολο να προστατευθεί. Οποιοδήποτε άτομο (όπως ο διαχειριστής συστήματος) με πρόσβαση σε ένα σύστημα παραγωγής μπορεί να κλέψει το μυστικό κλειδί και να το χρησιμοποιήσει για να αποκρυπτογραφήσει δεδομένα. Ενώ υπάρχουν κάποιες λύσεις υλικού, αυτό είναι ένα δύσκολο πρόβλημα αντιμετώπισης.

[12] Η συντομογραφία AES αναφέρεται στην αγγλική ορολογία Advanced Encryption Standard που χρησιμοποιείται και έχει υιοθετήσει η κυβέρνηση των ΗΠΑ για την προστασία διαβαθμισμένων πληροφοριών. Εφαρμόζεται σε υλικό και λογισμικό παγκοσμίως για την κρυπτογράφηση ευαίσθητων πληροφοριών. Κυρίως θεωρείται απαραίτητο εργαλείο για την ασφάλεια στον κυβερνοχώρο και των κυβερνητικών υπολογιστών.

1) Τα ασύμμετρα προγράμματα κρυπτογράφησης όπως το RSA, το DSA ή το ECC χρησιμοποιούν δύο κλειδιά: δημόσια και ιδιωτικά. Η κρυπτογράφηση δεδομένων χρησιμοποιεί το δημόσιο κλειδί και ακόμη και αν ένας εισβολέας επιτύχει αυτό το δημόσιο κλειδί, δεν είναι σε θέση να αποκρυπτογραφήσει προστατευμένα δεδομένα. Ωστόσο, η ασύμμετρη κρυπτογράφηση είναι πολύ πιο αργή από τη συμμετρική κρυπτογράφηση και, για το λόγο αυτό, σπάνια χρησιμοποιείται για κρυπτογράφηση δεδομένων από μόνη της.

2) Ομομορφική κρυπτογράφηση[13] (Homomorphic encryption)

Ο σκοπός της ομομορφικής κρυπτογράφησης είναι να επιτρέπει τον υπολογισμό των κρυπτογραφημένων δεδομένων. Η ομομορφική κρυπτογράφηση είναι μια μορφή κρυπτογράφησης που επιτρέπει τον υπολογισμό κρυπτογραφημένων δεδομένων, δημιουργώντας ένα κρυπτογραφημένο αποτέλεσμα το οποίο, όταν αποκρυπτογραφείται, ταιριάζει με το αποτέλεσμα των λειτουργιών σαν να είχαν πραγματοποιηθεί στα αρχικά δεδομένα.

2.3 Πραγματικό σενάριο εφαρμογής Γενικού Κανονισμού Προστασίας Δεδομένων

Στις προηγούμενες ενότητες μελετήσαμε τον κανονισμό, εντοπίζοντας τα κρίσιμα σημεία που τον διέπουν, αναλύσαμε, τα περιεχόμενά του και τις επιμέρους αυστηρές διαδικασίες και ενέργειες που απαιτούνται για την τήρησή του, τους επηρεαζόμενους φορείς του δημόσιου και ιδιωτικού τομέα σε συνάρτηση με τα φυσικά πρόσωπα που προσπαθεί να προστατεύσει η νέα αυτή νομοθεσία.
Στη συνέχεια, αναλύσαμε τις τεχνικές που η οδηγία προτρέπει να ακολουθηθούν από τους υπεύθυνους επεξεργασίας και εκτελούντες την επεξεργασία με σκοπό τη παράλληλη προστασία των φυσικών προσώπων στην επεξεργασία ευαίσθητων προσωπικών δεδομένων που τα αφορούν με την υπεύθυνη χρήση των νέων τεχνολογιών.

Σε αυτό το σημείο της εργασίας, θεωρείται χρήσιμη και κατάλληλη η μελέτη μιας πραγματικής χρήσης της εφαρμογής σε μια επιχείρηση του ιδιωτικού τομέα.

Για λόγους εχεμυθείας και εμπιστευτικότητας τα ονόματα των συνεργαζόμενων φορέων δε θα αναφερθούν, αλλά θα μελετηθεί από τεχνική πλευρά η εφαρμογή της οδηγίας και το κατά πόσο επηρεάζει τον τρόπο λειτουργίας και ανάπτυξης μιας επιχειρηματικής δραστηριότητας και θα εκτιμηθεί αν το μοντέλο που ακολουθήθηκε για την επίλυση του προβλήματος ήταν αποδοτικό και παραγωγικό.

---

[13] Κύρια λειτουργία της ομομορφικής κρυπτογράφησης είναι η εκτέλεση υπολογισμών σε κρυπτογραφημένα αρχεία χωρίς να χρειάζεται πρώτα η αποκρυπτογράφηση του συγκεκριμένου αρχείου.

Αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων προσωπικών δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας

Η προαναφερθείσα εταιρία, διατηρεί σε ένα σύστημα αρχειοθέτησης, συγκεκριμένα μέσω ενός Web Database Εξυπηρετητή και Oracle ως Σύστημα Διαχείρισης Βάσεων Δεδομένων, πληροφορίες και δεδομένα που αφορούν την παρακολούθηση αλιείας. Τα δεδομένα συλλέγονται ουσιαστικά μέσω ενός Vessel Monitoring System . Το συγκεκριμένο σύστημα συλλέγει δεδομένα ανά τακτά χρονικά διαστήματα μέσω συσκευών που έχουν τοποθετηθεί σε σκάφη που συνδέονται μέσω GPS με δορυφόρο για τη συλλογή των επιθυμητών πληροφοριών.

Παράλληλα, η ίδια εταιρία συμμετέχει σε ένα έργο με ερευνητικούς σκοπούς για την ανάλυση δεδομένων που αφορούν τα θαλάσσια ύδατα. Για το έργο αυτό είναι αναγκαίο η επεξεργασία μεγάλου όγκου δεδομένων συσχετιζόμενα με το περιεχόμενο πρόκειται να αναλυθεί. Ως ιδεατή λύση επιλέχθηκε η χρήση των δεδομένων που διατηρεί η εταιρία και προαναφερθήκαμε νωρίτερα σε τι αποσκοπούν.

Σε αυτό το σημείο της μελέτης και ανάλυσης του έργου, διαπιστώθηκε από τον Data Protection Officer της εταιρίας, ότι η χρησιμοποίηση των πληροφοριών από ένα άλλο project σε ένα άλλο, εφόσον, αφορά πρώτον επιχειρησιακό μοντέλο παρόλο που τα δεδομένα και οι πληροφορίες δε θα δημοσιευθούν και δεύτερον αν υπάρχουν μέσα σε αυτές τις πληροφορίες δεδομένα που μπορεί άμεσα ή έμμεσα με κάποιο συνδυασμό πληροφοριών και δεδομένων να οδηγήσουν στην ταυτοποίηση ενός φυσικού ατόμου, τότε, η εν λόγω δραστηριότητα, είναι αντίθετη στο νέο κανονισμό προστασίας φυσικών προσώπων έναντι της επεξεργασίας ευαίσθητων προσωπικών δεδομένων. Έτσι, ο DPO θεώρησε αναγκαίο για το σκοπό του ερευνητικού έργου, καταρχάς να γίνει έλεγχος αν υπάρχουν ευαίσθητες πληροφορίες που συλλέγονται μέσω του παραπάνω VMS και αν διαπιστωθεί ότι υπάρχουν να γίνει μετασχηματισμός της «στήλης» (column) που περιέχει την ευαίσθητη πληροφορία μέσω ενός αλγορίθμου SHA 256 bits, πριν την οποιαδήποτε ανάλυση και χρησιμοποίηση των δεδομένων αυτών, ώστε να διασφαλιστεί ότι η εταιρία δρα σύμφωνα με το GDPR.

Για να υλοποιηθεί αυτή η ανάθεση, αρχικά ήταν απαραίτητη η απομακρυσμένη σύνδεση στον εξυπηρετητή που υπήρχε εγκατεστημένο το σύστημα διαχείρισης βάσεων δεδομένων της Oracle. αυτό πραγματοποιήθηκε από το τοπικό σταθμό εργασίας μέσω του λογισμικού SQL Developer της Oracle που αποσκοπεί σε τέτοιες περιπτώσεις χρήσης.Στη συνέχεια διαπιστώθηκαν και δημιουργήθηκαν τα εξής θέματα προς επίλυση:

α) σε τη μορφή χρειάζεται να είναι τα νέα μετασχηματισμένα δεδομένα (csv[14], xlxs..)

β) αν χρειάζεται και απαιτείται η ενσωμάτωση στην Oracle Database του αλγορίθμου SHA 256 bits που ζητήθηκε.

---

[14] Το CSV είναι μορφή αποθήκευσης αρχείων, Comma Separated Values, που χρησιμοποιείται πολύ συχνά για την αποθήκευση και την επεξεργασία δεδομένων και θα το χρησιμοποιήσουμε στην υλοποίηση του αλγορίθμου.

γ) το πλήθος των δεδομένων και αν υπάρχουν όντως πληροφορίες που μπορεί να θεωρηθούν ευαίσθητες σύμφωνα με την οδηγία

δ) το αναμενόμενο κόστος και η χρήση της επεξεργαστικής μονάδας του εξυπηρετητή για να φέρει εις πέρας το μετασχηματισμό, την εντολή σε γλώσσα oracle sql που θα χρησιμοποιηθεί, πόσος όγκος δεδομένων απαιτείται για το ερευνητικό έργο και τέλος η δημιουργία νέας βάσης δεδομένων και νέων πινάκων , πιστά αντίγραφα των αρχικών πινάκων , δηλαδή με τους ίδιους τύπους δεδομένων για κάθε στήλη / column και ίδια ονόματα στηλών, αφού δεν προβλέπεται ο μετασχηματισμός για τη διατήρηση δεδομένων του υπάρχοντος έργου, παρά μόνο για τις ανάγκες του νέου ερευνητικού έργου.

Πώς επιλύθηκαν τα παραπάνω ζητήματα:

α) επιλέχθηκε η τελική μορφή των δεδομένων να εξάγονται σε .csv μορφή, που υλοποιείται εύκολα μέσω της Oracle Database.

β) λόγω της έκδοσης της Oracle Database, 9g, η οποία δεν παρέχει τον συγκεκριμένο αλγόριθμο, SHA 256 bits, ενσωματώθηκε στη βάση δεδομένων ο αναφερόμενος αλγόριθμος μέσω μιας διαδικασίας (procedure) που διαθέτει προς υλοποίηση η Oracle Database

γ) Αρχικά, εντοπίστηκε σε κάθε εγγραφή, ότι υπάρχει η στήλη IMEI, που αναφέρεται στον International Mobile Equipment Identity σε τύπο ακέραιου αριθμού και εν συνεπεία μπορεί να οδηγήσει με έμμεσο τρόπο και συνδυασμό άλλων πληροφοριών στην ταυτοποίηση φυσικού προσώπου. Οπότε, θεωρείται ευαίσθητο δεδομένο και απαιτείται ο μετασχηματισμός της στήλης σε όλες τις υπάρχουσες εγγραφές.

Στη συνέχεια, ζητήθηκε η μετατροπή αυτή για δεδομένα ενός μήνα που ανέρχονται σε 800000 περίπου εγγραφές, ενώ συνολικά για το νέο ερευνητικό έργο θα επρόκειτο να χρησιμοποιηθούν δεδομένα ενός έτους.

δ) Στο πρώτο μετασχηματισμό που πραγματοποιήθηκε για δεδομένα ενός μήνα, το κόστος επεξεργασίας και παράλληλα η απόδοση του εξυπηρετητή ήταν σε μεγάλο βαθμό σημαντικές, αφού για το πέρας εκτέλεσης της κατάλληλης SQL εντολής για εισαγωγή στο νέο πίνακα και ταυτόχρονα μετασχηματισμός της στήλης IMEI δαπανήθηκαν περίπου σαράντα πέντε λεπτά (45 minutes) , με αποτέλεσμα ο εξυπηρετητής να μη μπορεί να χρησιμοποιηθεί από άλλους χρήστες.

2.4 Συμπεράσματα πραγματικού σεναρίου εφαρμογής από την χειροκίνητη εφαρμογή χρήσης της οδηγίας:

Ο χειροκίνητος εντοπισμός και μετασχηματισμός πληροφοριών μέσα από βάσεις δεδομένων μεγάλης κλίμακας, αποτελεί σίγουρα, μέσα και από τα αποτελέσματα που καταγράφηκαν, μια νέα πρόκληση που πρέπει άμεσα να αντιμετωπιστεί, για να είναι εφικτή η ομαλή ανάπτυξη δραστηριοτήτων που σχετίζονται με την ανάλυση και επεξεργασία ευαίσθητων δεδομένων προσωπικού χαρακτήρα για ερευνητικούς, για επιχειρησιακούς και άλλους σκοπούς.

Μια προτεινόμενη λύση, θα ήταν να εντοπίζονται αυτόματα αυτές η πληροφορίες που είναι ευαίσθητες και να γίνει μια πρόβλεψη για μελλοντική χρήση τέτοιων τύπων δεδομένων, ώστε αυτόματα επίσης, να δημιουργούνται πιστά αντίγραφα βάσεων δεδομένων με τη νέα μετασχηματισμένη πληροφορία όπου, όταν και όποτε απαιτείται.

Για να γίνει αυτό εφικτό κρίνεται απαραίτητο η χρήση των τεχνολογιών και συγκεκριμένα η θεωρητική απόκτηση γνώσεων του επιστημονικού πεδίου της Πληροφορικής, Τεχνητή Νοημοσύνη, και εν συνεχεία η πρακτική υλοποίησή της μέσω της Μηχανικής Μάθησης, της Εξόρυξης Δεδομένων και μίας γλώσσας υψηλού επιπέδου ώστε να αναπτυχθεί επιτυχώς το αυτοματοποιημένο πληροφοριακό σύστημα ανίχνευσης ευαίσθητων δεδομένων σε βάσεις δεδομένων μεγάλης κλίμακας και στη συνέχεια ο επίσης προγραμματισμένος μετασχηματισμός και η εισαγωγή των αποτελεσμάτων σε νέα βάση δεδομένων ή πίνακα ή σύστημα αρχειοθέτησης για μελλοντική χρήση, λαμβάνοντας υπόψη και ενεργώντας με συναίσθηση και συνείδηση για τον νέο αυτό κανονισμό.

3.Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

3.1 Εισαγωγή στη θεωρία Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης

Η πρώτη απόπειρα ορισμού πραγματοποιήθηκε από τους Barr & Feigenbaum ως εξής:

«Η Τεχνητή Νοημοσύνη είναι ένας τομέας της επιστήμης των υπολογιστών που ασχολείται με τη σχεδίαση ευφυών υπολογιστικών συστημάτων, δηλαδή συστήματα που μπορούν να επιδείξουν χαρακτηριστικά που σχετίζονται με τη νοημοσύνη στην ανθρώπινη συμπεριφορά.»

Γενικά έχουν δημιουργηθεί πολλοί ορισμοί, χωρίς να είναι κάποιος απόλυτα ικανοποιητικός.

3.2 Μηχανική Μάθηση

Η **Μηχανική Μάθηση** είναι ένα πεδίο της **Τεχνητής Νοημοσύνης** που περιλαμβάνει την κατασκευή **αλγορίθμων αυτό - εκμάθησης** με σκοπό να αποκτήσει γνώση μέσα από **μεγάλο όγκο δεδομένων** ώστε να είναι σε θέση να κάνει **προβλέψεις**.

Θα μπορούσε αυτή η διαδικασία αντί να είναι αυτοματοποιημένη από κάποιο λογισμικό, να γίνεται χειροκίνητα από κάποιον αρμόδιο, αλλά λόγο της ραγδαίας ανάπτυξης της τεχνολογίας και του όγκου των δεδομένων που καθημερινά αυξάνεται με ταχύς ρυθμούς, πρακτικά είναι αδύνατο να επεξεργαστεί κάποιος τόσο μεγάλο όγκο δεδομένων σε συγκεκριμένο χρονικό περιθώριο.

Γι' αυτό το λόγο, η μηχανική μάθηση είναι μια αποτελεσματική **εναλλακτική προσέγγιση** της **απόκτησης της γνώσης** μέσα από δεδομένα που οδηγεί σε **βελτιστοποίηση του κόστους υπολογισμού**, την **πρόβλεψη συμβάντων** και την **λήψη αποφάσεων**.

Ορισμός:

Η μηχανική μάθηση είναι μια μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί την κατασκευή αναλυτικών μοντέλων.

Με χρήση αλγορίθμων που εκπαιδεύονται από επαναληπτικά δεδομένα, η μηχανική μάθηση επιτρέπει στους υπολογιστές να βρουν κρυφές πληροφορίες χωρίς να είναι ρητά προγραμματισμένοι για το πού να κοιτάξουν.

Πού ακριβώς μπορεί να χρησιμοποιηθεί η μηχανική μάθηση;

Παρακάτω παρουσιάζονται μερικά από τα πιο γνωστά πεδία εφαρμογής:

- Ανίχνευση απάτης.
- Αποτελέσματα αναζήτησης στο Web.
- Διαφημίσεις σε πραγματικό χρόνο σε ιστοσελίδες
- Βαθμολογία πίστωσης
- Πρόβλεψη αστοχιών εξοπλισμού
- Νέα μοντέλα τιμολόγησης
- Ανίχνευση εισβολής δικτύου
- Μηχανές προτάσεων
- Τμηματοποίηση πελατών

- Ανάλυση συναισθημάτων κειμένου
- Περικοπή πελατών
- Φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου.

Ποια είναι η διαφορά μεταξύ μηχανικής μάθησης και νευρωνικών δικτύων; (Neural Networks)

Τα νευρωνικά δίκτυα είναι ένας τρόπος μοντελοποίησης βιολογικών συστημάτων νευρώνων με μαθηματικά.

- Αυτά τα δίκτυα μπορούν στη συνέχεια να χρησιμοποιηθούν για την επίλυση εργασιών που πολλοί άλλοι τύποι αλγορίθμων δεν μπορούν (π.χ. ταξινόμηση εικόνας).
- Η βαθιά μάθηση (deep learning) αναφέρεται απλά σε νευρωνικά δίκτυα με περισσότερα από ένα κρυμμένα επίπεδα.

Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, υπάρχουν διαφορετικές μεθοδολογίες εκμάθησης ενός πληροφοριακού συστήματος. Εμείς θα αναφερθούμε σε δύο μεγάλες κατηγορίες:

1) Εποπτευόμενη μάθηση

2) Μη εποπτευόμενη μάθηση

Καταλήγουμε ότι:

Α) Η μηχανική μάθηση χρησιμοποιεί αυτοματοποιημένα μαθηματικά μοντέλα.

Β) Τα νευρωνικά δίκτυα είναι ένας τύπος αρχιτεκτονικής μηχανικής μάθησης που διαμορφώθηκε σύμφωνα με βιολογικούς νευρώνες.

Γ) Η λεγόμενη βαθιά μάθηση είναι ένα νευρωνικό δίκτυο με περισσότερα από ένα κρυφό στρώμα (layer).

Αφού κάναμε μια μικρή αναφορά στις διαφορετικές κατηγορίες ώστε να μπορούμε να κατανοήσουμε τη διαφορά τους, μπορούμε να επικεντρωθούμε στην εποπτευόμενη εκμάθηση ενός συστήματος, η οποία θα είναι και η διαδικασία που θα ακολουθήσουμε για την ανάπτυξη και αξιολόγηση του συστήματός μας, με στόχο την ανίχνευση ευαίσθητων προσωπικών δεδομένων, ώστε να χρησιμοποιηθεί στην επίλυση ενός επίκαιρου θέματος που δημιουργήθηκε με την εφαρμογή του νέου ΓΚΠΔ.

Η μηχανική μάθηση χωρίζεται σε τρεις (3) διαφορετικές κατηγορίες:

i) Supervised Learning
ii) Reinforcement Learning
iii) Unsupervised Learning

i) Με την **εποπτευόμενη μάθηση** μπορούμε να κάνουμε προβλέψεις.

Κύριος στόχος της εποπτευόμενης μάθησης είναι η **μάθηση από δεδομένα** που μας επιτρέπουν να κάνουμε **προβλέψεις** για τα **άγνωστα** ή μελλοντικά δεδομένα.

Με τον όρο «εποπτευόμενη» εννοούμε πρακτικά, ένα **σύνολο δειγμάτων** όπου τα επιθυμητά αποτελέσματα είναι **εξ' αρχής γνωστά**. Έτσι, μπορούμε να **εκπαιδεύσουμε ένα μοντέλο** χρησιμοποιώντας έναν "***supervised machine learning algorithm***" σύμφωνα πάντα με τα δεδομένα που έχουμε στη διάθεσή μας, τις ανάγκες και τα επιθυμητά αποτελέσματα που μας ενδιαφέρουν.

Αυτό υλοποιείται χρησιμοποιώντας:

- **Ετικέτες κλάσης** (class labels) – **Εργασία ταξινόμησης** (classification task)
- **Οπισθοδρόμηση** (regression)

Labels → Training Data → Machine Learning Algorithm → New Data → Predictive Model → Prediction

__Ταξινόμηση__ για την πρόβλεψη __ετικετών κλάσεων__

Αυτή η διαδικασία υπάγεται ως μια υποκατηγορία του Supervised Learning που περιλαμβάνει τις **ετικέτες των κλάσεων** (class labels) με στόχο την **ομαδοποίηση των περιπτώσεων** και/ή τη **δυαδική ταξινόμηση**.

Το **μοντέλο πρόβλεψης** μπορεί να εκπαιδευτεί μέσω ενός **supervised learning algorithm** για την ταξινόμηση **πολλαπλών κατηγοριών**.

Χρήση της __οπισθοδρόμησης__ (regression) για την πρόβλεψη συνεχών αποτελεσμάτων.

Η πρόβλεψη συνεχών αποτελεσμάτων υλοποιείται μέσω της διαδικασίας της **ανάλυσης της οπισθοδρόμησης** (regression analysis)

Η ανάλυση της οπισθοδρόμησης μπορεί να εκφραστεί μέσα από μεταβλητές:

Α) **predictor** (explanatory) **variables**

B) **continuous response variable** (outcome)

Ένα είδος της ανάλυσης οπισθοδρόμησης είναι **η γραμμική οπισθοδρόμηση** όπου έχουμε μια μεταβλητή **Predictor x** και μια μεταβλητή **Response y**.

ii) Reinforcement Learning

Η **μη εποπτευόμενη μάθηση** σχετίζεται άμεσα με την **εποπτευόμενη μάθηση**.

Στόχος της είναι η ανάπτυξη ενός συστήματος (agent) που θα **βελτιστοποιεί** την **απόδοση** βασισμένη σε **αλληλεπιδράσεις** με το **περιβάλλον**.

Πρακτικά, μέσω της μη εποπτευόμενης μάθησης, δίνεται η δυνατότητα να **ανακαλυφθούν δομές δεδομένων** από **άγνωστους** τύπους δεδομένων.

Η υλοποίηση αυτή πραγματοποιείται μέσω της εξερεύνησης δομές δεδομένων ώστε να παράγουμε δεδομένα για τα οποία δε γνωρίζουμε εξ' αρχής το τύπο μεταβλητών ή όπως αποκαλείται της ***reward function***.

Ένα σημαντικό πρόβλημα της **μηχανικής μάθησης** είναι ο **όγκος** των **δεδομένων** (Big Data) που απαιτείται για να εκπαιδευτεί το «έξυπνο» λογισμικό και πως θα μπορέσουμε να αναλύσουμε αυτό τον όγκο δεδομένο ώστε να μπορέσουμε να βγάλουμε **χρήσιμα συμπεράσματα** μέσα από αυτά.

Η πιο συνήθης και **αποτελεσματική προσέγγιση** αυτού του προβλήματος είναι η **ταξινόμηση των δεδομένων** σε **ομάδες/συστάδες** μέσω αλγορίθμων **συσταδοποίησης / ομαδοποίησης** (clustering) μεγάλου όγκου δεδομένων.

Ένας αλγόριθμος **clustering** που πλέον χρησιμοποιείται συχνά είναι ο K – Means.

Πρακτικά, μέσω της συσταδοποίησης των δεδομένων, δίνει τη δυνατότητα στο σύστημα να ταξινομεί τα δεδομένα που έχουν κοινά στοιχεία (πχ τύπο δεδομένων ή είναι ιδιότητες του ίδιου αντικειμένου). Μετά το clustering θα έχουμε οργανώσει τα δεδομένα σε υποομάδες (subgroups), δημιουργώντας παράλληλα ένα βαθμό «συγγένειας» των δεδομένων.

Η **συσταδοποίηση οδηγεί** σε **εποπτευόμενη ταξινόμηση** των δεδομένων.

Ένα εξίσου σημαντικό πρόβλημα κατά την επεξεργασία των δεδομένων είναι **διαστάσεις** που έχουν τα **δεδομένα** (dimensionality). Είναι λογικό να σκεφτεί κανείς

την **πολυπλοκότητα** που θα έχει η **επεξεργασία** των δεδομένων αν αυτά είναι **τρισδιάστατα** ή παραπάνω.

Ένας τρόπος αντιμετώπισης του ζητήματος αυτού είναι η **μείωση** (reduction) των **διαστάσεων** των **δεδομένων** με σκοπό τη μείωση του **κόστους υπολογισμού**. Αυτό υλοποιείται με μαθηματικές γνώσεις της **Γραμμικής Άλγεβρας**, που μας δίνει τη δυνατότητα να επεξεργαστούμε ένα **Πίνακα (Matrix) N x M διαστάσεων** και το **μετασχηματίσουμε** σε μικρότερες διαστάσεις, **διατηρώντας** τις **σχετιζόμενες πληροφορίες**.

Στη μηχανική μάθηση υιοθετούμε και προσαρμόζουμε το γνωστικό αντικείμενο της Γραμμικής Άλγεβρας σύμφωνα με τις ανάγκες μας. Για το λόγο αυτό θεωρούμε ένα πίνακα (matrix) και ως δείκτες:

i → ιωστό δείγμα προς εξάσκηση

j →jωστή διάσταση του συνόλου δεδομένων.

3.3 Εποπτευόμενη μηχανική μάθηση

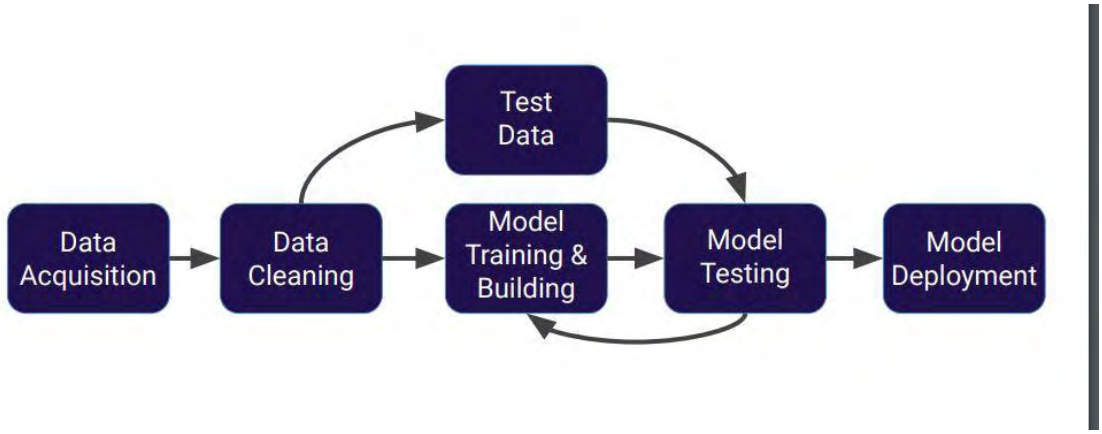
Σε αυτή την κατηγορία οι αλγόριθμοι εκπαιδεύονται με επίβλεψη, χρησιμοποιώντας επισημασμένα παραδείγματα, όπως μια εισαγωγή όπου είναι γνωστή η επιθυμητή έξοδος.

Για παράδειγμα, ένα τμήμα κειμένου θα μπορούσε να έχει μια ετικέτα κατηγορίας, όπως:

1) Spam εναντίον (vs) νόμιμου email Review
2) Θετική εναντίον (vs) αρνητική κριτική ταινίας

Το σύστημα λαμβάνει ένα σύνολο εισόδων μαζί με τις αντίστοιχες σωστές εξόδους και με αυτή τη μέθοδο ο αλγόριθμος μαθαίνει συγκρίνοντας την πραγματική του έξοδο με σωστά αποτελέσματα για την εύρεση σφαλμάτων. Στη συνέχεια τροποποιείται ανάλογα και το μοντέλο. Η εποπτευόμενη μάθηση χρησιμοποιείται συνήθως στο εφαρμογές όπου προβλέπονται πιθανά μελλοντικά γεγονότα βάσει των διαθέσιμων δεδομένων.

Ακολουθεί ένα διάγραμμα με τα βήματα και τη συσχέτιση των βημάτων που ακολουθούνται στην εποπτευόμενη μάθηση:



Εικόνα 3.3.1 Διάγραμμα βημάτων εποπτευόμενης μάθησης

Το πρώτο βήμα - η απόκτηση των δεδομένων - μπορεί να φαίνεται ως μια απλή διαδικασία εύρεσης κατάλληλου συνόλου δεδομένων (dataset), αλλά αυτό δεν είναι απόλυτα ακριβές. Αρχικά, πρέπει να έχουμε εμπεδώσει ακριβώς ποιο είναι το πρόβλημα που καλούμαστε να επιλύσουμε, να γνωρίζουμε τι είδους πληροφορίες αναζητάμε και τέλος να είναι επαρκές σε όγκο. Όπως έχουμε αναφέρει και προηγουμένως, όσο μεγαλύτερος και έγκυρος είναι ο όγκος των πληροφοριών θα έχουμε και έγκυρα αποτελέσματα. Αυτό ισχύει σε μεγάλο βαθμό στην κατηγορία της εποπτευόμενης μάθησης, αφού οι πληροφορίες που ήδη έχουμε στην κατοχή μας καθορίζουν και το επιθυμητό αποτέλεσμα. Αν έχουμε άστοχες ή ψευδείς πληροφορίες τότε θα είναι ανάλογο και το αποτέλεσμα της εκμάθησης.

Το δεύτερο βήμα - ο καθαρισμός των δεδομένων - και η προεπεξεργασία των δεδομένων είναι ίσως το πιο κομβικό βήμα στο οποίο παρουσιάζονται και τα περισσότερα λάθη. Για να είμαστε σε θέση να εφαρμόσουμε τον αλγόριθμο εκμάθησης, τα δεδομένα πρέπει να είναι σε τέτοια μορφή που το μοντέλο να μπορεί να τα κατανοεί και να είναι μορφοποιημένα στις ανάγκες μας.

Για παράδειγμα, συχνά στην εποπτευόμενη μάθηση χρησιμοποιούμε Regression - οπισθοδρόμηση - που είναι ένα μαθηματικό μοντέλο. Τα κατηγορικά, ή αλλιώς μη αριθμητικά δεδομένα, δεν μπορούν να ληφθούν σε αυτό το μοντέλο γιατί απλά δεν μπορεί να τα καταλάβει ο αλγόριθμος. Για το βήμα της προεπεξεργασίας των δεδομένων, θα αναφερθούμε αναλυτικά κατά την υλοποίηση του βήματος αυτού, σε επόμενο κεφάλαιο. Προς το παρόν σημαντικό είναι να αναφερθεί ότι χωρίς τη βιβλιοθήκη της Pandas που διαθέτει η Python, τότε αυτό το βήμα θα ήταν πρακτικά πολύ δύσκολο να υλοποιηθεί.

Τα επόμενα βήματα που ακολουθούν την προεπεξεργασία, δηλαδή το μοντέλο εκπαίδευσης, της αξιολόγησης και των αποτελεσμάτων υλοποιούνται με τη βοήθεια

της Sklearn βιβλιοθήκης στην Python, που έχει δημιουργηθεί αποκλειστικά για αυτό το σκοπό.

Αυτό που μόλις αναφέραμε είναι μια απλοποιημένη προσέγγιση στην εποπτευόμενη μάθηση, συνεπάγεται, όμως, ένα μεγάλο ζήτημα: είναι ορθό να χρησιμοποιούμε το μεμονωμένο διαχωρισμό των δεδομένων αξιολογώντας την απόδοση των μοντέλων μας;

Αφού επεξηγήσαμε απλά τη διαδικασία, θα εξηγήσουμε πως είναι δυνατό να ενημερώνουμε παραμέτρους του μοντέλου επαναληπτικά, όποτε θεωρούμε πως είναι αναγκαίο.

Για να επιλύσουμε αυτό το ζήτημα, τα δεδομένα χωρίζονται συχνά σε τρία (3) σύνολα:

1) Δεδομένα εκπαίδευσης → Χρησιμοποιούνται για την εκπαίδευση παραμέτρων μοντέλου.
2) Δεδομένα επικύρωσης → Χρησιμοποιούνται για τον προσδιορισμό του μοντέλου υπερπαραμέτρων, για εναρμόνιση του μοντέλου.
3) Δεδομένα δοκιμής → Χρησιμοποιούνται για τη λήψη μέτρησης της τελικής απόδοσης.

Αυτό σημαίνει ότι, αφού δούμε τα αποτελέσματα στο τελικό δοκιμαστικό σύνολο, δεν μπορούμε να επιστρέψουμε και να προσαρμόσουμε κανένα παράμετρο του μοντέλου!

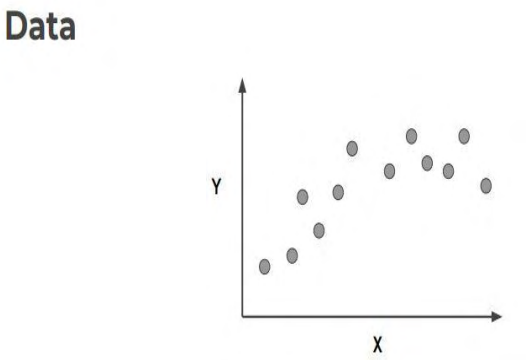
● Αυτό το τελικό μέτρο είναι αυτό που ονομάζουμε (labeling) αληθινή απόδοση του μοντέλου.

Τώρα που κατανοούμε την πλήρη διαδικασία για την εποπτευόμενη μάθηση, ας δούμε τα σημαντικά θέματα της υπερβολικής και της ανεπαρκής τοποθέτησης.

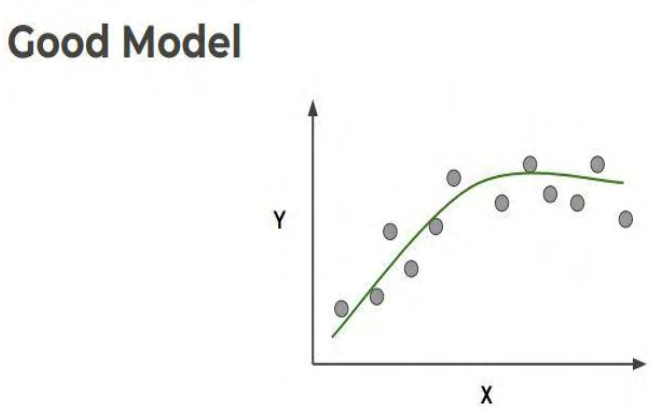
Υπερβολική τοποθέτηση (Overfitting)

Το μοντέλο ταιριάζει πάρα πολύ με το θόρυβο από τα δεδομένα. Αυτό συχνά οδηγεί σε χαμηλά σφάλματα κατά την εκπαίδευση του συνόλου, αλλά παράλληλα σε υψηλά ποσοστά σφαλμάτων στα σετ δοκιμής / επικύρωσης.

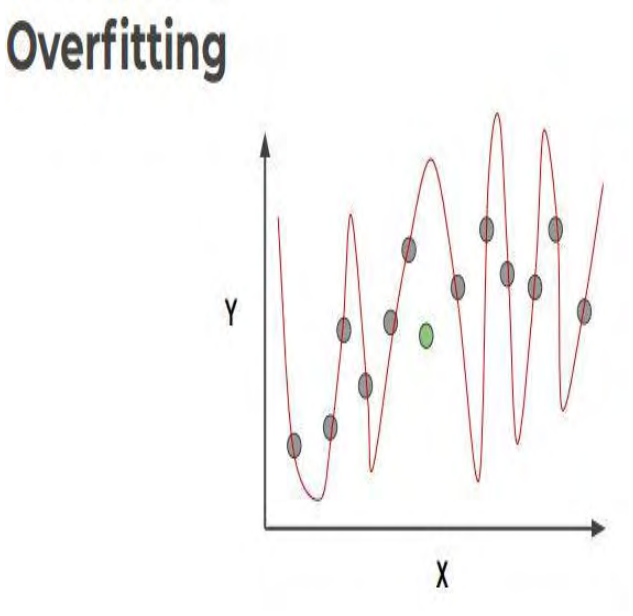
Παρακάτω ακολουθούν διαγράμματα με τις διαφορές μεταξύ ενός καλού μοντέλου και ενός μοντέλου με Overfitting.



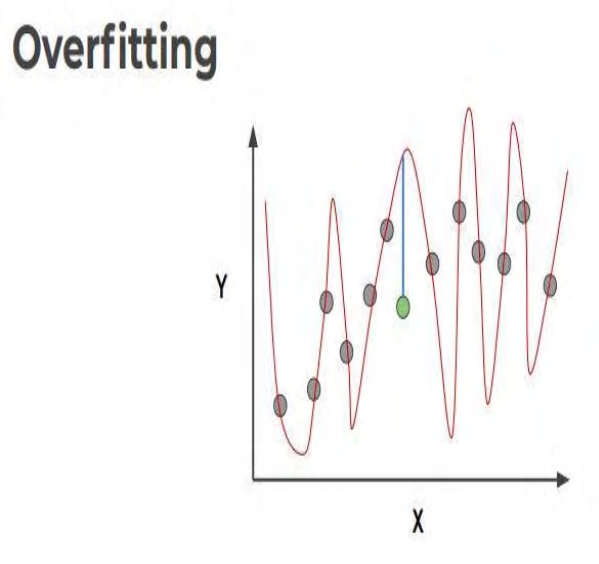
Εικόνα 3.3.2 Αναπαράσταση Δεδομένων



Εικόνα 3.3.3 Αναπαράσταση καλού μοντέλου



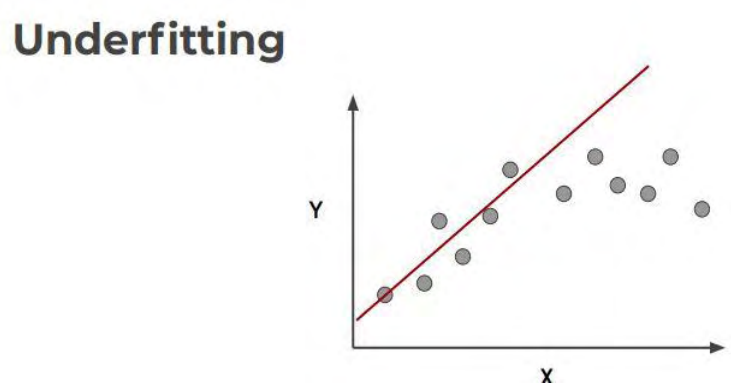
Εικόνα 3.3.4 Απεικόνιση υπερβολικής Τοποθέτησης



Εικόνα 3.3.5 Γράφημα Overfitting

Ανεπαρκή Τοποθέτηση (Underfitting)

Το μοντέλο δεν συλλαμβάνει την υποκειμενική τάση των δεδομένων και δεν ταιριάζει με τα δεδομένα αρκετά καλά. Αποτέλεσμα είναι η χαμηλή διακύμανση αλλά και η υψηλή μεροληψία. Συχνά διαπιστώνεται σε υπερβολικά απλά μοντέλα.



Εικόνα 3.3.6 Αναπαράσταση Underfitting

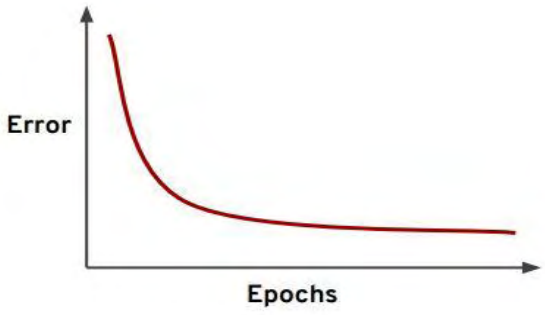
Αυτά τα δεδομένα ήταν εύκολο να οπτικοποιηθούν, αλλά πώς μπορούμε να διαπιστώσουμε το underfitting και το overfitting όταν ασχολούμαστε με πολυδιάστατα σύνολα δεδομένων;

Ας υποθέσουμε πρώτα ότι εκπαιδεύσαμε ένα μοντέλο και μετά ας μετρήσουμε το σφάλμα του κατά τη διάρκεια της εκμάθησης.

Όταν σκεφτόμαστε την υπερβολική τοποθέτηση και την ανεπαρκή τοποθέτηση θέλουμε να έχουμε κατά νου τη σχέση της απόδοσης του μοντέλου στο training set έναντι του test / evaluation set.

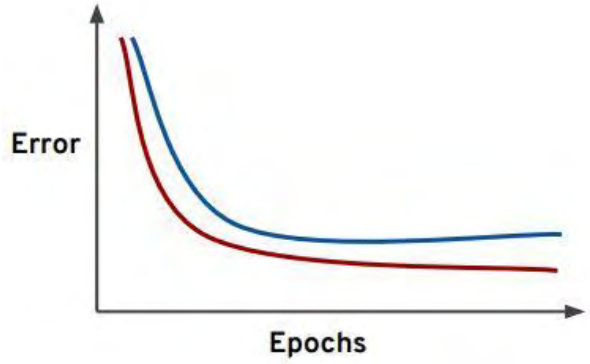
Ας υποθέσουμε ότι χωρίζουμε τα δεδομένα μας σε train set και σε test set.

Πρώτα βλέπουμε την απόδοση στο training set:



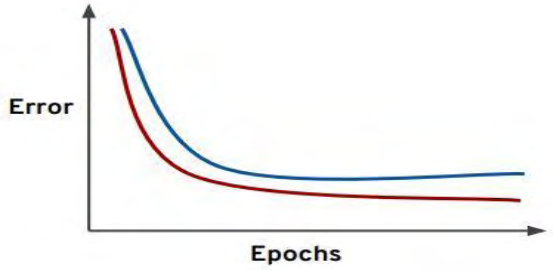
Εικόνα 3.3.7 Απόδοση Training Set

Στη συνέχεια ελέγχουμε την απόδοση στο test set:



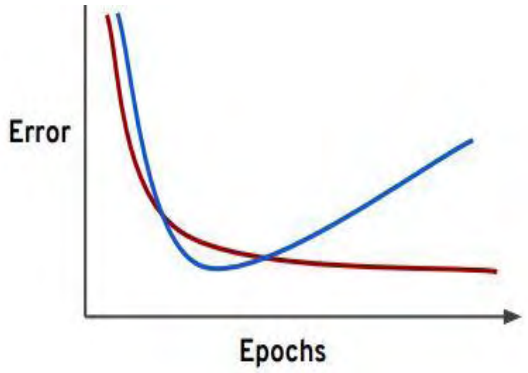
Εικόνα 3.3.8 Απόδοση Test Set

Στην ιδανική περίπτωση, το μοντέλο θα είχε καλή απόδοση και στα δύο με παρόμοια συμπεριφορά, όπως φαίνεται στην εικόνα:



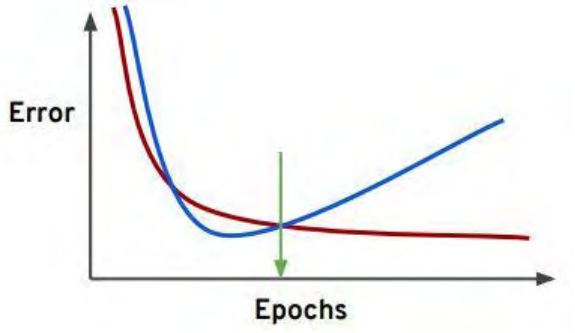
Εικόνα 3.3.9 Αναπαράσταση αποδόσεων Training και Test Sets - Καλή απόδοση

Αλλά τι θα συμβεί αν έχουμε υπερβολική τοποθέτηση (overfitting) στο training set δεδομένων; Αυτό σημαίνει ότι θα έχουμε καλή απόδοση στο νέο test set δεδομένων!



3.3.10 Αναπαράσταση Απόδοσης με Overfitting

Αυτή είναι μια καλή ένδειξη ότι υπερβάλλουμε σχετικά με το training set και θα πρέπει να αναζητήσουμε το σημείο για να σταματήσουμε το χρόνο του training set!



Εικόνα 3.3.11 Αναπαράσταση σημείου Υπερβολικής Τοποθέτησης

Θα ελέγξουμε ξανά αυτήν την ιδέα όταν θα αρχίσουμε να δημιουργούμε το ζητούμενο μοντέλο εκπαίδευσης.

3.4 Αξιολόγηση μοντέλου

Μόλις είδαμε ότι, μόλις η μαθησιακή διαδικασία έχει ολοκληρωθεί, θα χρησιμοποιήσουμε μετρήσεις απόδοσης για να αξιολογήσουμε αν το μοντέλο μας είναι αποτελεσματικό και αποδοτικό.

Ας εξηγήσουμε τις μετρήσεις ταξινόμησης (classification metrics) με περισσότερη λεπτομέρεια.

Συνήθως, σε οποιαδήποτε εργασία ταξινόμησης (classification task), το μοντέλο μπορεί να έχει μόνο δύο αποτελέσματα: το μοντέλο ήταν σωστό στην πρόβλεψη ή το μοντέλο ήταν λανθασμένο στην πρόβλεψη.

Ευτυχώς, το λανθασμένο έναντι του σωστού επεκτείνεται σε καταστάσεις όπου έχουμε πολλές κατηγορίες.

Για τους σκοπούς της εξήγησης των μετρήσεων, ας φανταστούμε μια δυαδική ταξινόμηση (binary classification), όπου έχουμε μόνο δύο διαθέσιμες κλάσεις.

Στο δικό μας πρόβλημα, θα προσπαθήσουμε να προβλέψουμε εάν μια πληροφορία είναι ευαίσθητη πληροφορία και μπορεί να οδηγήσει στην ταυτοποίηση φυσικού ατόμου ή όχι.

Δεδομένου ότι εκτελούμε εποπτευόμενη μάθηση, θα κάνουμε μια πρώτη εφαρμογή κατάρτιση ενός μοντέλου για δεδομένα εκπαίδευσης και, στη συνέχεια, θα δοκιμάσουμε το μοντέλο σε δεδομένα δοκιμών.

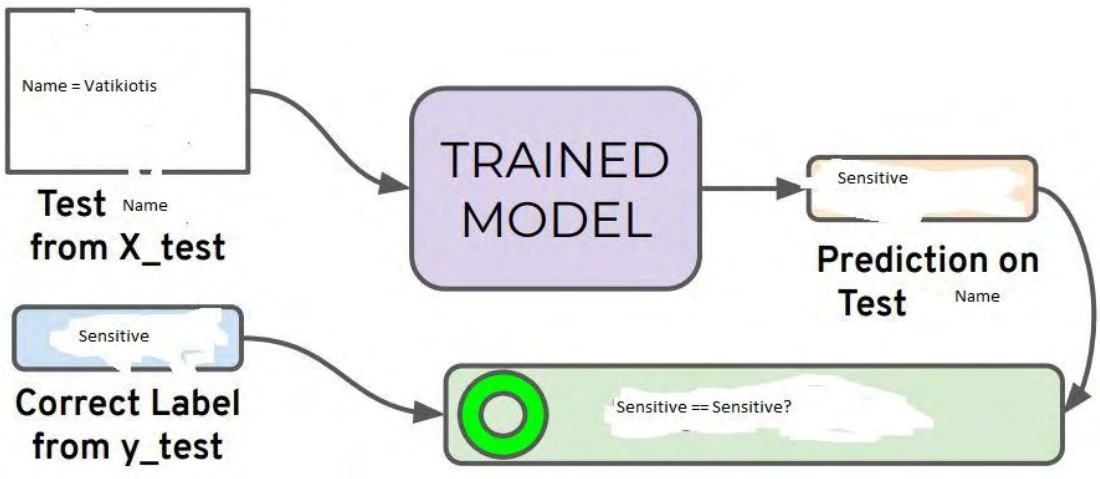
Μόλις έχουμε τις προβλέψεις του μοντέλου από τα δεδομένα (X_test), τα συγκρίνουμε με τις αληθινές τιμές y (οι σωστές τιμές).

Ας υποθέσουμε ότι έχουμε μια πληροφορία (στήλη) με ετικέτα (label) Name και ότι το περιεχόμενο της ετικέτας Name ισούται με (=) Vatikiotis. Εφόσον έχουμε εκπαιδεύσει ήδη το μοντέλο μας με αυτή την πληροφορία, τότε έχουμε:

Name = Vatikiotis → Trained Model

Test Name from X_test → Sensitive → Correct label from y_test.

Έτσι πρέπει να συγκρίνουμε το αποτέλεσμα του X_test με αυτό του y_test. Αν έχουμε ίδια αποτελέσματα τότε έχουμε κάνει σωστή πρόβλεψη, όπως φαίνεται και σχηματικά:



Εικόνα 3.4.1 Απεικόνιση Παραδείγματος Μοντέλου Γραμμικής Οπισθοδρόμησης

Στο τέλος θα έχουμε μια σωστή μέτρηση προβλέψεων και πλήθος λανθασμένων προβλέψεων.

Πρέπει βασικά να συνειδητοποιήσουμε ότι στον πραγματικό κόσμο δεν είναι όλες οι προβλέψεις λάθος και ότι οι σωστές προβλέψεις επίσης δεν έχουν ίση αξία!

Επίσης, στον πραγματικό κόσμο, μόνο μια μετρική δεν αρκεί για να επαληθεύσεις ή να απορρίψεις μια πρόβλεψη. Θα μπορούσαμε να οργανώσουμε τις προβλεπόμενες τιμές μας σε σύγκριση με τις πραγματικές τιμές σε ένα πίνακα που ονομάζεται Confusion Matrix.

Ακρίβεια: Η ακρίβεια στα προβλήματα ταξινόμησης είναι ο αριθμός σωστών προβλέψεων από το μοντέλο διαιρούμενο με το συνολικό αριθμό προβλέψεων.

Για παράδειγμα, εάν το σύνολο X_test ήταν 100 στήλες και το μοντέλο μας πρόβλεψε σωστά 80 στήλες, για τον είναι ευαίσθητη πληροφορία ή όχι, τότε έχουμε 80/100.

0,8 ή 80% ακρίβεια.

Η ακρίβεια είναι χρήσιμη όταν στοχεύουμε σε τάξεις (classes) που είναι καλά ισορροπημένες. Στο παράδειγμά μας, θα έχουμε περίπου την ίδια ποσότητα στηλών που δεν είναι ευαίσθητη πληροφορία με αυτές που είναι ευαίσθητη πληροφορία.

Η ακρίβεια δεν είναι καλή επιλογή σε ισορροπημένες τάξεις (classes). Φανταστείτε ότι είχαμε 99 στήλες που ήταν ευαίσθητη πληροφορία και 1 στήλη που δεν ήταν. Εάν το μοντέλο μας ήταν απλώς μια γραμμή, πάντα προβλέποντας την ευαίσθητη πληροφορία, τότε θα είχαμε 99% ακρίβεια, κάτι που δεν επαληθεύει και δεν προσδιορίζει το σωστό αποτέλεσμα.

Σε τέτοιες περιπτώσεις πρέπει να ανατρέξουμε και να κατανοήσουμε τις έννοιες της ανάκλησης (recall) και της ακρίβειας (precision).

Ανάκληση (recall):

Είναι η ικανότητα ενός μοντέλου να βρει όλες τις σχετικές περιπτώσεις σε ένα σύνολο δεδομένων.

Ο ακριβής ορισμός της ανάκλησης είναι ο αριθμός πραγματικών θετικών, διαιρεμένος με τον αριθμό των πραγματικών θετικών συν τον αριθμό των ψευδών αρνητικών.

Ακρίβεια (precision):

Είναι η ικανότητα ενός μοντέλου ταξινόμησης να προσδιορίσει μόνο τα σχετικά σημεία δεδομένων. Η ακρίβεια ορίζεται ως ο αριθμός πραγματικών θετικών, διαιρεμένος με τον αριθμό των πραγματικών θετικών συν τον αριθμό των ψευδών θετικών.

Ανάκληση και ακρίβεια

Συχνά συναντάμε μια ανταλλαγή μεταξύ ανάκλησης και ακρίβειας. Ενώ η ανάκληση εκφράζει την ικανότητα εύρεσης όλων των σχετικών σε ένα σύνολο δεδομένων, η ακρίβεια εκφράζει την αναλογία των σημείων δεδομένων που το μοντέλο μας λέει ότι ήταν σχετικά και στην πραγματικότητα ήταν σχετικά.

F1 – Score:

Σε περιπτώσεις όπου θέλουμε να βρούμε ένα βέλτιστο συνδυασμό ακρίβειας και ανάκλησης, μπορούμε να συγκρίνουμε τις δυο αυτές μετρικές χρησιμοποιώντας το F1 - Score.

Το F1 - Score είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης λαμβάνοντας και τις δύο μετρικές στην ακόλουθη εξίσωση:

$$F1 = 2 * \frac{ακρίβεια * ανάκληση}{ακρίβεια + ανάκληση}$$

Χρησιμοποιούμε τον αρμονικό μέσο όρο αντί για έναν απλό μέσο όρο γιατί συνυπολογίζει τις ακραίες τιμές. Ένας ταξινομητής με ακρίβεια 1,0 και ανάκληση 0,0 έχει έναν απλό μέσο όρο 0,5 αλλά F1 - score 0.

Πίνακας Σύγχυσης – Confusion Matrix :

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**

Εικόνα 3.4.2 Πίνακας Σύγχυσης – Confusion Matrix

		predicted condition		
	total population	prediction positive	prediction negative	Prevalence = Σ condition positive / Σ total population
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection = Σ TP / Σ condition positive
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm = Σ FP / Σ condition negative
	Accuracy = (Σ TP + Σ TN) / Σ total population	Positive Predictive Value (PPV), Precision = Σ TP / Σ prediction positive	False Omission Rate (FOR) = Σ FN / Σ prediction negative	Positive Likelihood Ratio (LR+) = TPR / FPR
		False Discovery Rate (FDR) = Σ FP / Σ prediction positive	Negative Predictive Value (NPV) = Σ TN / Σ prediction negative	Negative Likelihood Ratio (LR−) = FNR / TNR

Εικόνα 3.4.3 Πίνακας Σύγχυσης με Εξισώσεις Σύγκρισης

Το κύριο σημείο που πρέπει να θυμόμαστε με τον πίνακα σύγχυσης και τις διάφορες υπολογισμένες μετρικές είναι ότι είναι όλες βασικοί τρόποι σύγκρισης των προβλεπόμενων τιμών έναντι των πραγματικών τιμών.

Το τι συνιστά «καλές» μετρικές, θα εξαρτηθεί πραγματικά από τη συγκεκριμένη κατάσταση!

3.5 Οπισθοδρόμηση και εργαλεία υλοποίησης

3.5.1 Θεωρία οπισθοδρόμησης

Ας περάσουμε τώρα να συζητήσουμε την αξιολόγηση μοντέλων οπισθοδρόμησης. Η οπισθοδρόμηση είναι μια εργασία που εκτελείται όταν ένα μοντέλο προσπαθεί να προβλέψει συνεχείς τιμές (continuous values) σε αντίθεση με τις κατηγορικές τιμές, που είναι ταξινόμηση (classification).

Σε αυτή την περίπτωση, οι μετρικές ακρίβειας και ανάκλησης που μελετήσαμε προηγουμένως δεν είναι χρήσιμες για την οπισθοδρόμηση αφού χρειαζόμαστε μετρικές σχεδιασμένες για συνεχείς τιμές.

Για παράδειγμα, απόπειρα πρόβλεψης του κόστους ενός σπιτιού δεδομένου ότι γνωρίζουμε τα χαρακτηριστικά του, είναι μια εργασία οπισθοδρόμησης (regression

task). Ενώ αντίθετα η απόπειρα πρόβλεψης της χώρας του σπιτιού μέσα από τα χαρακτηριστικά του θα ήταν μια εργασία ταξινόμησης (classification task).

Σε αυτό το σημείο θα συζητήσουμε τις βασικότερες μετρικές αξιολόγησης της οπισθοδρόμησης:

Μέσο απόλυτο σφάλμα (Mean Absolute Error – MAE)

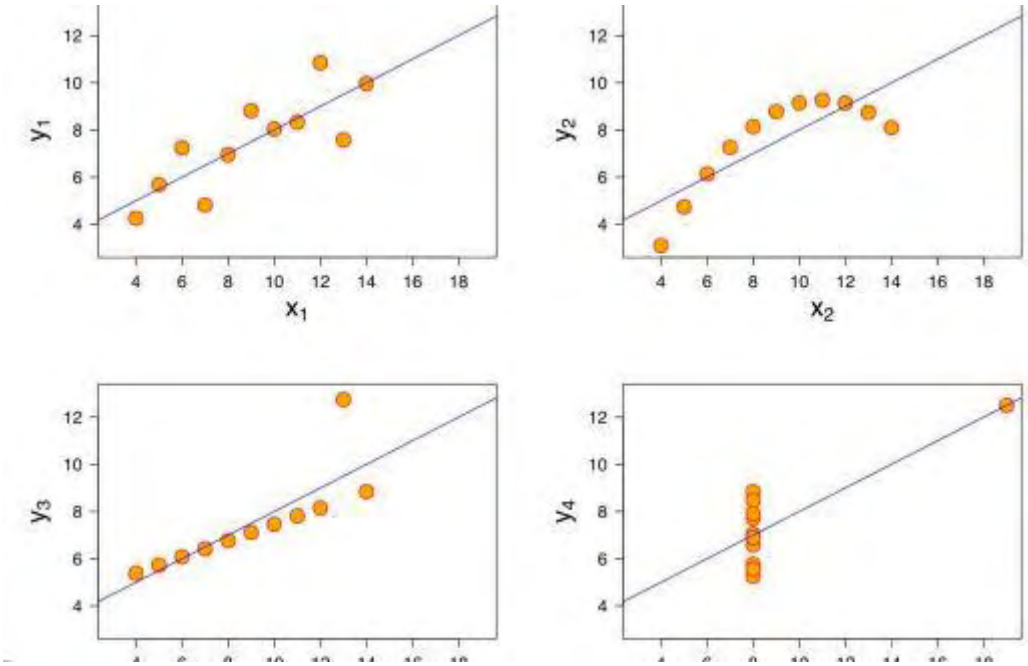
Μέσο σφάλμα τετραγώνου (Mean Squared Error – MSE)

Σφάλμα ρίζας μέσου όρου τετραγώνου (Root Mean Square Error - RMSE)

1) **Μέσο απόλυτο σφάλμα (MAE):** Αυτός είναι ο μέσος όρος της απόλυτης τιμής λαθών:

$$\frac{1}{n}\sum_{i}^{n} = 1 \mid y_i - \hat{y}_i \mid$$

Ωστόσο, το MAE δεν θα συνυπολογίσει τα μεγάλα σφάλματα:



Εικόνα 3.5.1.1 Αναπαράσταση Αποτελεσμάτων MAE

Σφάλμα μέσου τετραγώνου (MSE): Αυτός είναι ο μέσος όρος του τετραγώνου σφαλμάτων. Μεγαλύτερα σφάλματα σημειώνονται περισσότερο από με το MAE, καθιστώντας το MSE πιο δημοφιλές:

$$\frac{1}{n}\sum_{i}^{n} = 1\,(\,y_i - \,\hat{y}_i\,)^2$$

Σφάλμα ρίζας μέσου όρου τετραγώνου (RMSE): Αυτή είναι η ρίζα του μέσου όρου του τετράγωνου σφαλμάτων. Πιο δημοφιλής (έχει τις ίδιες μονάδες με το y):

$$\sqrt{\frac{1}{n}\sum_{i}^{n} = 1\,(\,y_i - \,\hat{y}_i\,)^2}$$

Μέχρι αυτό το σημείο μελετήσαμε όλες τις μετρικές για την αξιολόγηση της πρόβλεψης και του μοντέλου μας ανάλογα με το αν έχουμε κατηγορηματικές ή αριθμητικές τιμές. Στη συνέχεια της εργασίας, θα υλοποιήσουμε το στάδιο της συλλογής, προεπεξεργασίας και εκτέλεσης του αλγορίθμου εποπτευόμενης μάθησης χρησιμοποιώντας τις τεχνικές της οπισθοδρόμησης και της ταξινόμησης όπου απαιτείται, ώστε να δούμε τι αποτέλεσμα έχει εν τέλει το εγχείρημά μας. Στόχος εξ' αρχής είναι η αντιμετώπιση του ΓΚΠΔ μέσα από τις νέες τεχνολογίες και συγκεκριμένα μέσω της Python και της θεωρίας της μηχανικής μάθησης που αναλύσαμε παραπάνω.

3.5.2 Εργαλεία υλοποίησης

Για να γίνει πράξη η υλοποίηση ενός συστήματος που θα λειτουργεί ως βοηθητικό εργαλείο για τους Υπεύθυνους Επεξεργασίας και τους εκτελούντες την επεξεργασία ευαίσθητων προσωπικών δεδομένων χρειαζόμαστε τα εξής:

1) Μια γλώσσα προγραμματισμού υψηλού επιπέδου, κατάλληλη για το project και την ανάπτυξη αλγορίθμου μηχανικής μάθησης.

Η πιο κοινή και γνωστή γλώσσα προγραμματισμού που χρησιμοποιείται ευρέως σε προβλήματα Machine Learning και Data Science είναι η Python, αφού διαθέτει πολλαπλές βιβλιοθήκες (modules - libraries) που βοηθούν σημαντικά στην υλοποίηση και δημιουργία αλγορίθμων σχετικά με το πρόβλημά μας.

Μερικές από τις σημαντικότερες βιβλιοθήκες τις Python που θα χρησιμοποιήσουμε για το project είναι:

Pandas → Σημαντική βιβλιοθήκη που περιέχει συναρτήσεις για την επεξεργασία και τη διαμόρφωση ενός συνόλου δεδομένου πχ DataFrame, με εύκολο και αποτελεσματικό τρόπο.

Matplotlib → Βιβλιοθήκη που βοηθά σημαντικά στη διαγραμματική απεικόνιση των πληροφοριών, πριν, κατά τη διάρκεια και στο πέρας του αλγορίθμου μηχανικής μάθησης.

Sklearn → Μια εξαιρετικά χρήσιμη βιβλιοθήκη που περιέχει αλγορίθμους και τα πιο γνωστά μοντέλα μηχανικής μάθησης όπως: KNN, Bayesian algorithms, clustering, regression κλπ.

2) Πολύ σημαντικοί παράγοντες για την εκτέλεση και την υλοποίηση μηχανικής μάθησης είναι η εύρεση μεγάλου όγκου δεδομένων και πληροφοριών όπως και, παράλληλα, πληροφορίες έγκυρες και κατάλληλες σχετικές με θέματα ευαίσθητων προσωπικών δεδομένων.

Λαμβάνοντας υπόψη τα παραπάνω, ανατρέχουμε στο πρώτο βήμα για την εκτέλεση και περάτωση της εργασίας, που δεν είναι άλλο από τη συγκέντρωση κατάλληλων πληροφοριών σε συνδυασμό με το μεγάλο όγκο που απαιτείται. Επίσης, θα εκτελέσουμε αλγόριθμο μηχανικής μάθησης με τη μέθοδο της εποπτευόμενης εκπαίδευσης (Supervised Machine Learning Algorithm), όπου το κλειδί και το κύριο κομμάτι για επιτυχή αποτελέσματα είναι η συγκέντρωση πληροφοριών. Όσες περισσότερες οι πληροφορίες που μπορούμε να έχουμε στη διάθεσή μας τόσο μεγαλύτερες είναι οι πιθανότητες επιτυχίας. Προτού μιλήσουμε για τα δεδομένα, πρέπει να αναφέρουμε τα βήματα που ακολουθούνται για την ανάπτυξη και την αξιολόγηση ενός αλγορίθμου μηχανικής μάθησης.

3.6 Σχετικές δημοσιεύσεις και άρθρα σχετικά με την ανίχνευση προσωπικών πληροφοριών

Με βάση τα όσα προαναφέρθηκαν για την προστασία των προσωπικών δεδομένων και τους τρόπους ανίχνευσής τους , ανατρέξαμε σε πηγές και δημοσιεύσεις ερευνητών με προσεγγίσεις στην δική μας λογική, δηλαδή την ανάπτυξη πληροφοριακού συστήματος για την αυτοματοποιημένη ανίχνευση ευαίσθητων προσωπικών δεδομένων.

Με βάση τις οδηγίες του κανονισμού η δημοσίευση εστιάζει σε βασικές μεθοδολογίες:

Α) Διεξαγωγή ανάλυσης κινδύνων σε δεδομένα και πληροφορίες που προέρχονται από βάσεις δεδομένων, σύνολα δεδομένων και έγγραφα για την κατανόηση των πληροφοριών που εμπίπτουν στο GDPR.

Β) Κατηγοριοποίηση ή ταξινόμηση κάθε πληροφορίας που θεωρείται ευαίσθητη ως “προσωπική πληροφορία – sensitive”

Γ) Επεξεργασία των προσωπικών πληροφοριών με βάση την οδηγία GDPR.

Επίσης αναφέρει τους τρόπους και τις τεχνικές ψευδοανωνυμοποίησης και ανωνυμοποίησης που προτείνει ο ΓΚΠΔ για τον μετασχηματισμό των δεδομένων.

Στη συνέχεια επικεντρώνεται στην υιοθέτηση τεχνικών μηχανικής μάθησης για την ανίχνευση προσωπικών δεδομένων και την ενσωμάτωση αλγορίθμου ψευδωνυμοποίησης ή ανωνυμοποίησης για την διευκόλυνση επεξεργασίας στους φορείς / οργανισμούς.

Η διαδικασία και ο αλγόριθμος που προτείνεται φαίνεται παρακάτω σχηματικά:

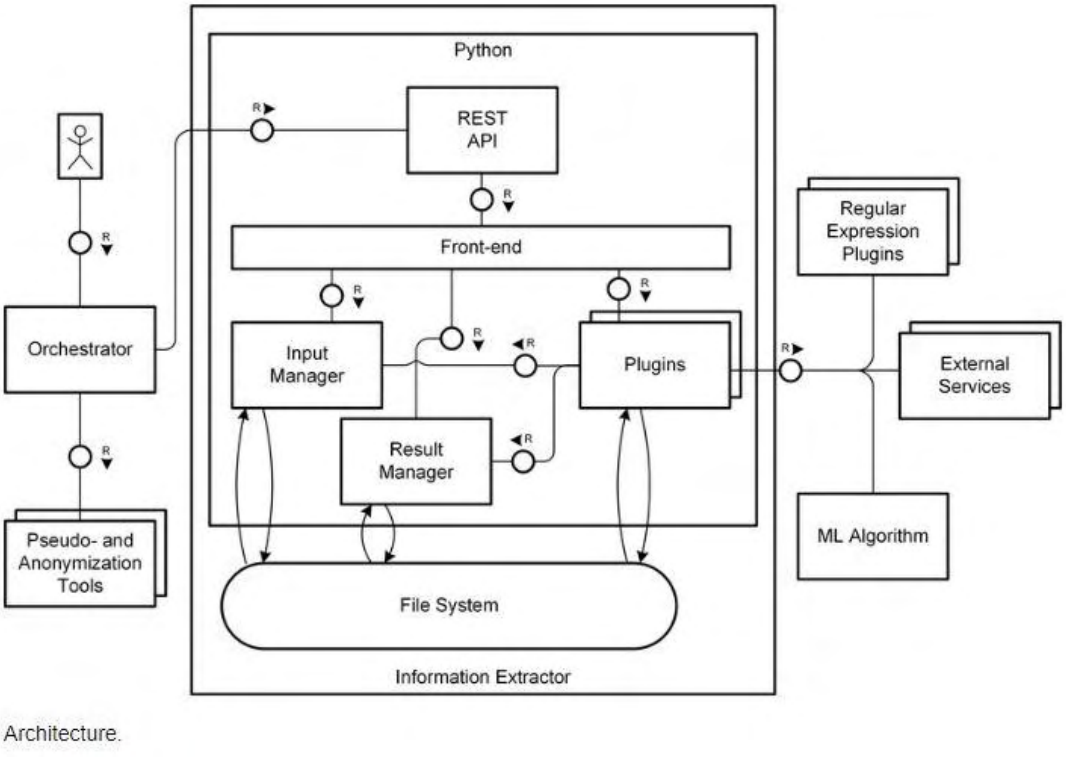
Algorithm 1: Process

Data: A generic *input_text*, in any format
Result: An *output_text* with *input_text* contents but anonymized personal information
begin initialization:
  train a machine learning algorithm with personal data training sets;
end
begin execution:
  `// step R2`
  evaluate *input_text* against the machine learning algorithm previously trained to identify personal information;
  `// step R2a`
  pseudo- or anonymize personal information;
  create *output_text* from *input_text* inserting anonymized personal information;
  return *output_text*;
end

Εικόνα 3.6.1 Απεικόνιση Αλγορίθμου Σύμφωνα με το άρθρο

Το άρθρο αυτό αναφέρεται στο GDPR, αναλύει τις οδηγίες και τους κανόνες και προτείνει τη Μηχανική Μάθηση για την εναρμόνιση στη νέα αυτή οδηγία, γεγονός που ταιριάζει με το δικό μας σκεπτικό.

Τέλος παρουσιάζεται στο άρθρο αυτό η αρχιτεκτονική του πληροφοριακού συστήματος που απαιτείται:



Εικόνα 3.6.2 Αρχιτεκτονική Πληροφοριακού Συστήματος

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) http://goo.gl/LfwxGe

[2] McCallum A, Li W (2003) Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lex-icons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics, pp 188–191

[3] Ohm P (2009) Broken promises of privacy: Responding to the surprising failure of anonymization. Ucla L Rev 57:1701

[4] Towards Personal Data Identification and Anonymization Using Machine Learning Techniques: ADBIS 2018 Short Papers and Workshops, AI*QA, BIGPMED, CSACDB, M2U, Big Data MAPS, ISTREND, DC, Budapest, Hungary, September, 2-5, 2018, Proceedings

4. Υλοποίηση τεχνικού μέρους εργασίας

4.1 Εύρεση πληροφοριών και προεπεξεργασία δεδομένων

Για το πρώτο βήμα εφαρμογής του αλγορίθμου απαιτείται η συγκέντρωση πληροφοριών σχετικά με τον ΓΚΠΔ. Λόγω της φύσης των πληροφοριών είναι δύσκολο η ανάκτηση μεγάλου όγκου πληροφοριών από τον παγκόσμιο ιστό.

Έστω, λοιπόν ότι έχουμε στην κατοχή ένα αρχείο .csv που αφορά πελατολόγιο μιας τράπεζας έχοντας κάποιες πληροφορίες σχετικά με τους πελάτες[15]. Θα ξεκινήσουμε με την Python και με τη βοήθεια του Jupyter Notebook ως γραφικό περιβάλλον και θα δούμε βήμα βήμα τη διαδικασία προεπεξεργασίας δεδομένων.

Αρχικά, εισάγουμε στο κώδικά μας, τις απαραίτητες βιβλιοθήκες:

Import the relevant libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
from sklearn.linear_model import LinearRegression
```

Εικόνα 4.1.1 Εισαγωγή βιβλιοθηκών

Με τη βοήθεια της Pandas διαβάζουμε το σύνολο δεδομένων που έχουμε με τη συνάρτηση read_csv που διαθέτει και για να δούμε ότι έχει γίνει σωστή εισαγωγή δεδομένων εκτυπώνουμε τις πρώτες πέντε (5) εγγραφές παρατηρώντας παράλληλα και τι είδους δεδομένα έχουμε στις στήλες του αρχείου:

```
raw_data = pd.read_csv('bank-full.csv',sep=';')
#print the first five rows to see the data
raw_data.head()
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown

Εικόνα 4.1.2 Φόρτωση δεδομένων από αρχείο .csv

[15] Το σύνολο δεδομένων σχετίζεται με ενέργειες άμεσου μάρκετινγκ ενός πορτογαλικού τραπεζικού ιδρύματος. Οι διαδικασίες μάρκετινγκ βασίστηκαν σε τηλεφωνικές κλήσεις. Συχνά, χρειάζονταν περισσότερες από μία επαφές στον ίδιο πελάτη. Εμείς αντλούμε αυτές τις πληροφορίες θέλοντας να αναδείξουμε αν κάποιες από αυτές τις στήλες θεωρούνται ευαίσθητη πληροφορία η όχι.

Παραπάνω βλέπουμε πώς έχουμε κατηγορηματικές και αριθμητικές στήλες. Σε πρώτη φάση θα δούμε πως μπορούμε να διαχειριστούμε τις αριθμητικές τιμές. Επίσης, με μια πρώτη ματιά μπορούμε να δούμε κάποια χαρακτηριστικά που μπορεί να ταυτοποιήσουν σε συνδυασμό με άλλες πληροφορίες ένα φυσικό άτομο. Τέτοιες στήλες μπορεί να είναι η ηλικία (age), η ημέρα (day) και ο μήνας (month). Επειδή όμως οι συγκεκριμένες ημερομηνίες δεν αφορούν ημερομηνία γέννησης δεν είναι σωστό να χαρακτηρισθούν ως ευαίσθητες. Εδώ συναντάμε και το πρώτο εμπόδιο. Αν θεωρήσουμε ότι κάθε χαρακτηριστικό με την ονομασία day, month, year, age είναι ευαίσθητες πληροφορίες τότε το εγχείρημά μας και το σύστημα θα εκπαιδευτεί με βάσει λανθασμένες προσεγγίσεις. Έτσι, για να επιλυθεί αυτό το ζήτημα θα βγάλουμε μετρικές για κάθε στήλη ξεχωριστά, σκεπτόμενοι πως κάθε πληροφορία κρύβει γνωρίσματα που είναι μοναδικά και με βάση αυτά, δηλαδή από κάποια γνωρίσματα φύσεως κάθε πληροφορίας θα μπορούμε να διακρίνουμε ορθά αν μια πληροφορία είναι ευαίσθητη και όχι απλά βλέποντας μια ονομασία γιατί το νόημα και η σημασία κάθε πληροφορίας μπορεί να αλλάζει σύμφωνα με την πηγή των πληροφοριών.

Ας ασχοληθούμε πρώτα με την προεπεξεργασία των αριθμητικών χαρακτηριστικών του συνόλου δεδομένων που διαθέτουμε. Η στήλη ‘age[16]’ είναι μια αριθμητική μεταβλητή που αφορά την ηλικία των πελατών μιας τράπεζας. Κάθε DataFrame στη Python έχει τη συνάρτηση describe() που υπολογίζει και εμφανίζει το πλήθος, το μέγιστο, το ελάχιστο, το μέσο όρο και άλλα κρυφά στατιστικά γνωρίσματα αριθμητικών πληροφοριών που με γυμνό μάτι δε μπορούμε να διακρίνουμε. Έτσι έχουμε:

```
#first we create a list having some metrics from describe function
#such as 25%, 50%, 75%, standard deviation, min, max , mean
AGE = [raw_data['age'].describe()]
```

```
AGE
```

```
[count    45211.000000
 mean        40.936210
 std         10.618762
 min         18.000000
 25%         33.000000
 50%         39.000000
 75%         48.000000
 max         95.000000
 Name: age, dtype: float64]
```

Εικόνα 4.1.3 Αποθήκευση γνωρισμάτων στήλης σε μορφή Πίνακα

[16] Όπου: **mean** → μέσος όρος, **std** → τυπική απόκλιση, **min** → ελάχιστο, **max** → μέγιστο, **median** → η μεσαία τιμή του πλήθους τιμών σε ταξινομημένο σύνολο, **range** → εύρος τιμών, **variance** → διακύμανση τιμών, **mode** → η τιμή που εμφανίζεται πιο συχνά σε ένα σύνολο τιμών δεδομένων. Αυτές θα είναι και οι μετρικές όσο αφορά τις αριθμητικές μεταβλητές που θα χρησιμοποιήσουμε ως είσοδο για τον αλγόριθμο Γραμμικής Οπισθοδρόμησης.

Τώρα, διαχωρίζουμε τα γνωρίσματα της στήλης age σε μία λίστα. Όμως, στόχος μας είναι να δημιουργήσουμε ένα νέο DataFrame[17] έχοντας ως στήλες τα γνωρίσματα και ως εγγραφή τη στήλη από το αρχικό DataFrame. Έτσι, με τη βοήθεια της Python και της Pandas είναι σχετικά εύκολο να μετασχηματίσουμε τα δεδομένα μας, στη μορφή που επιθυμούμε. Οπότε μετατρέπουμε τη λίστα σε DataFrame και έχουμε:

```
new_age = pd.DataFrame(AGE)
```

```
new_age
```

	count	mean	std	min	25%	50%	75%	max
age	45211.0	40.93621	10.618762	18.0	33.0	39.0	48.0	95.0

Εικόνα 4.1.4 Μετατροπή Πίνακα σε DataFrame

Επιπλέον, λόγω της ανάγκης περισσότερων στηλών – γνωρισμάτων από τα οποία θα βγουν τα features για το μοντέλο μας, δηλαδή τα κρίσιμα χαρακτηριστικά που θα κρίνουν μια πληροφορία ως ευαίσθητη ή όχι μπορούμε με τη βοήθεια της βιβλιοθήκης statistics που διαθέτει η Python να προσθέσουμε median, range και variance στις στήλες:

```
#lets import statics module to add more metrics
import statistics as st
#add the mode, median and rage to our metrics using append method
new_age['median'] = st.median(raw_data['age'])
new_age['range'] = new_age['max'] - new_age['min']
#next we will put variance as a metric-field
new_age['variance'] = st.variance(raw_data['age'])
new_age
```

Εικόνα 4.1.5 Προσθήκη γνωρισμάτων στο DataFrame

Στο τέλος της επεξεργασίας της στήλης age, θα έχουμε δημιουργήσει μια εγγραφή με ιδιότητες και χαρακτηριστικά που θα χρησιμοποιήσουμε στο μοντέλο μας:

	count	mean	std	min	25%	50%	75%	max	median	range	variance
age	45211.0	40.93621	10.618762	18.0	33.0	39.0	48.0	95.0	39	77.0	112.758107

Εικόνα 4.1.6 Απεικόνιση Αριθμητικών γνωρισμάτων

[17] Το DataFrame είναι μια δισδιάστατη δομή δεδομένων διαφορετικών τύπων που χρησιμοποιείται ως αντικείμενο της Pandas και θα το χρησιμοποιήσουμε για τη συλλογή και αποθήκευση των μετρικών για την εποπτευόμενη μάθηση.

Όπως φαίνεται παραπάνω, όλες αυτές οι στήλες θα αποτελούν τις ανεξάρτητες μεταβλητές X1, X2, X3,...,Xn. Την εξαρτημένη μεταβλητή μας y (sensitive), θα την δημιουργήσουμε εμείς σε μια κλίμακα από το 0 έως το 1 εκτιμώντας κατά πόσο η πληροφορία είναι ευαίσθητη. Έτσι θα έχουμε τις κατάλληλες πληροφορίες για να εκπαιδεύσουμε το σύστημά μας με έναν αλγόριθμο εποπτευόμενης μάθησης (supervised machine learning algorithm), αφού θα γνωρίζουμε εξ' αρχής όλες τις ανεξάρτητες αλλά παράλληλα και την εξαρτημένη – ζητούμενη μεταβλητή.

Για αυτό το λόγο θεωρούμε πως η ηλικία πελατών μιας τράπεζας θεωρείται 0.7 ή 70% ως ευαίσθητη πληροφορία αφού σε συνδυασμό με άλλη μπορεί να προδώσει την ταυτότητα ενός ατόμου. Έτσι έχουμε:

	count	mean	std	min	25%	50%	75%	max	median	range	variance	mode	sensitive
age	45211.0	40.93621	10.618762	18.0	33.0	39.0	48.0	95.0	39	77.0	112.758107	32	0.7

Εικόνα 4.1.7 Προσθήκη εξαρτημένης μεταβλητής - sensitive

Αν επαναλάβουμε τη διαδικασία για κάθε αριθμητική μεταβλητή που έχουμε στο συγκεκριμένο σύνολο δεδομένων δημιουργείται το εξής αποτέλεσμα:

	count	mean	std	min	25%	50%	75%	max	median	range	variance	mode	sensitive
age	45211.0	40.936210	10.618762	18.0	33.0	39.0	48.0	95.0	39	77.0	1.127581e+02	32	0.7
balance	45211.0	1362.272058	3044.765829	-8019.0	72.0	448.0	1428.0	102127.0	448	110146.0	9.270599e+06	0	0.3
day	45211.0	15.806419	8.322476	1.0	8.0	16.0	21.0	31.0	16	30.0	6.926361e+01	20	0.4
duration	45211.0	258.163080	257.527812	0.0	103.0	180.0	319.0	4918.0	180	4918.0	6.632057e+04	124	0.2
campaign	45211.0	2.763841	3.098021	1.0	1.0	2.0	3.0	63.0	2	62.0	9.597733e+00	1	0.2
pdays	45211.0	40.197828	100.128746	-1.0	-1.0	-1.0	-1.0	871.0	-1	872.0	1.002577e+04	-1	0.2
previous	45211.0	0.580323	2.303441	0.0	0.0	0.0	0.0	275.0	0	275.0	5.305841e+00	0	0.2

Εικόνα 4.1.8 Απεικόνιση συνόλου για την Γραμμική Οπισθοδρόμηση

Βλέπουμε πως από μια πηγή δεδομένων, δημιουργήσαμε 8 εγγραφές με στήλες διαμορφωμένες από την επεξεργασία τους για το μοντέλο μας. Αυτό σημαίνει ότι υπάρχει η αδυναμία εύρεσης μεγάλου όγκου δεδομένων (big data) και πρέπει να κάνουμε συσχετίσεις ή να προσθέσουμε στο ίδιο Data Frame που δημιουργήσαμε επιπλέον δεδομένα και εγγραφές από άλλες πηγές. Στην παρούσα φάση, θα είχαμε under fitting στην εκτέλεση της εκμάθησης και το αποτέλεσμα δε θα ήταν αντιπροσωπευτικό. Πέρα από αυτό το γεγονός που καλούμαστε να αντιμετωπίσουμε, δεν έχουμε συμπεριλάβει τις κατηγορηματικές μεταβλητές και δεν έχουμε αναφέρει τι είδους χαρακτηριστικά γνωρίσματα μπορούμε να αντλήσουμε από αυτές ώστε να συμπεριληφθούν αρμονικά στο Data Frame που δημιουργήσαμε.

Σημαντική παρατήρηση σε αυτό το σημείο είναι ότι εφόσον επιθυμούμε να εφαρμόσουμε τη μεθοδολογία της οπισθοδρόμησης πρέπει όλες οι μεταβλητές να είναι αριθμητικές οπότε θα πρέπει να διαμορφώσουμε και να επεξεργαστούμε κατάλληλα τις κατηγορηματικές μεταβλητές ώστε να μπορέσουμε να εφαρμόσουμε στην πράξη την οπισθοδρόμηση (regression).

Όπως αναφέραμε προηγουμένως, πρέπει να εισάγουμε περισσότερο όγκο πληροφοριών ώστε το μοντέλο μας να είναι αποτελεσματικό. Σε αυτό το σημείο θα εισάγουμε και θα επεξεργαστούμε παρόμοια ένα σύνολο δεδομένων που αφορά την επίδοση φοιτητών στο μάθημα των μαθηματικών. Παρόλο που τα σύνολα δεδομένων δεν έχουν καμία σχέση μεταξύ τους, εμάς μας αφορά να βγάλουμε συμπεράσματα βάση των μοναδικών ιδιοτήτων κάθε στήλης όπως έχουμε προαναφέρει, οπότε δεν μας ενοχλεί η πηγή και το θέμα του συνόλου (dataset). Μετά από κατάλληλη επεξεργασία δεδομένων όπως έχουμε δει και σε προηγούμενο βήμα, έχουμε ως αποτέλεσμα ένα σύνολο δεδομένων που από 8 αυξήθηκε σε 20 εγγραφές:

	25%	50%	75%	count	max	mean	median	min	mode	range	sensitive	std	variance
0	33.0	39.0	48.0	45211.0	95.0	40.936210	39	18.0	32	77.0	0.7	10.618762	1.127581e+02
1	72.0	448.0	1428.0	45211.0	102127.0	1362.272058	448	-8019.0	0	110146.0	0.3	3044.765829	9.270599e+06
2	8.0	16.0	21.0	45211.0	31.0	15.806419	16	1.0	20	30.0	0.4	8.322476	6.926361e+01
3	103.0	180.0	319.0	45211.0	4918.0	258.163080	180	0.0	124	4918.0	0.2	257.527812	6.632057e+04
4	1.0	2.0	3.0	45211.0	63.0	2.763841	2	1.0	1	62.0	0.2	3.098021	9.597733e+00
5	-1.0	-1.0	-1.0	45211.0	871.0	40.197828	-1	-1.0	-1	872.0	0.2	100.128746	1.002577e+04
6	0.0	0.0	0.0	45211.0	275.0	0.580323	0	0.0	0	275.0	0.2	2.303441	5.305841e+00
7	16.0	17.0	18.0	395.0	22.0	16.696203	17	15.0	16	7.0	0.7	1.276043	1.628285e+00
8	2.0	3.0	4.0	395.0	4.0	2.749367	3	0.0	4	4.0	0.4	1.094735	1.198445e+00
9	2.0	2.0	3.0	395.0	4.0	2.521519	2	0.0	2	4.0	0.4	1.088201	1.184180e+00
10	1.0	1.0	2.0	395.0	4.0	1.448101	1	1.0	1	3.0	0.2	0.697505	4.865129e-01
11	1.0	2.0	2.0	395.0	4.0	2.035443	2	1.0	2	3.0	0.2	0.839240	7.043244e-01
12	0.0	0.0	0.0	395.0	3.0	0.334177	0	0.0	0	3.0	0.2	0.743651	5.530168e-01
13	4.0	4.0	5.0	395.0	5.0	3.944304	4	1.0	4	4.0	0.2	0.896659	8.039967e-01
14	3.0	3.0	4.0	395.0	5.0	3.235443	3	1.0	3	4.0	0.2	0.998862	9.977254e-01
15	2.0	3.0	4.0	395.0	5.0	3.108861	3	1.0	3	4.0	0.2	1.113278	1.239388e+00
16	1.0	1.0	2.0	395.0	5.0	1.481013	1	1.0	1	4.0	0.2	0.890741	7.934203e-01
17	1.0	2.0	3.0	395.0	5.0	2.291139	2	1.0	1	4.0	0.2	1.287897	1.658678e+00
18	3.0	4.0	5.0	395.0	5.0	3.554430	4	1.0	5	4.0	0.2	1.390303	1.932944e+00
19	0.0	4.0	8.0	395.0	75.0	5.708861	4	0.0	0	75.0	0.2	8.003096	6.404954e+01

Εικόνα 4.1.9 Προσθήκη νέων πληροφοριών στο DataFrame

Συμπεριλαμβάνει τα αποτελέσματα επεξεργασίας δεδομένων που αφορούν πελατολόγιο τράπεζας και επίδοσης φοιτητών στα μαθηματικά. Μπορούμε να τα διαχωρίσουμε βλέποντας πως το 1ο dataset είχε 45211 εγγραφές αρχικά, ενώ το 2ο dataset είχε αρχικά 395.

Στο σύνολο δεδομένων μας, προσθέτουμε και τις επεξεργασμένες πληροφορίες που αφορούν επιδόσεις φοιτητών στο μάθημα της πορτογαλικής γλώσσας:

	25%	50%	75%	count	max	mean	median	min	mode	range	sensitive	std	variance
0	33.0	39.0	48.0	45211.0	95.0	40.936210	39	18.0	32	77.0	0.7	10.618762	1.127581e+02
1	72.0	448.0	1428.0	45211.0	102127.0	1362.272058	448	-8019.0	0	110146.0	0.3	3044.765829	9.270599e+06
2	8.0	16.0	21.0	45211.0	31.0	15.806419	16	1.0	20	30.0	0.4	8.322476	6.926361e+01
3	103.0	180.0	319.0	45211.0	4918.0	258.163080	180	0.0	124	4918.0	0.2	257.527812	6.632057e+04
4	1.0	2.0	3.0	45211.0	63.0	2.763841	2	1.0	1	62.0	0.2	3.098021	9.597733e+00
5	-1.0	-1.0	-1.0	45211.0	871.0	40.197828	-1	-1.0	-1	872.0	0.2	100.128746	1.002577e+04
6	0.0	0.0	0.0	45211.0	275.0	0.580323	0	0.0	0	275.0	0.2	2.303441	5.305841e+00
7	16.0	17.0	18.0	395.0	22.0	16.696203	17	15.0	16	7.0	0.7	1.276043	1.628285e+00
8	2.0	3.0	4.0	395.0	4.0	2.749367	3	0.0	4	4.0	0.4	1.094735	1.198445e+00
9	2.0	2.0	3.0	395.0	4.0	2.521519	2	0.0	2	4.0	0.4	1.088201	1.184180e+00
10	1.0	1.0	2.0	395.0	4.0	1.448101	1	1.0	1	3.0	0.2	0.697505	4.865129e-01
11	1.0	2.0	2.0	395.0	4.0	2.035443	2	1.0	2	3.0	0.2	0.839240	7.043244e-01
12	0.0	0.0	0.0	395.0	3.0	0.334177	0	0.0	0	3.0	0.2	0.743651	5.530168e-01
13	4.0	4.0	5.0	395.0	5.0	3.944304	4	1.0	4	4.0	0.2	0.896659	8.039967e-01
14	3.0	3.0	4.0	395.0	5.0	3.235443	3	1.0	3	4.0	0.2	0.998862	9.977254e-01
15	2.0	3.0	4.0	395.0	5.0	3.108861	3	1.0	3	4.0	0.2	1.113278	1.239388e+00
16	1.0	1.0	2.0	395.0	5.0	1.481013	1	1.0	1	4.0	0.2	0.890741	7.934203e-01
17	1.0	2.0	3.0	395.0	5.0	2.291139	2	1.0	1	4.0	0.2	1.287897	1.658678e+00
18	3.0	4.0	5.0	395.0	5.0	3.554430	4	1.0	5	4.0	0.2	1.390303	1.932944e+00
19	0.0	4.0	8.0	395.0	75.0	5.708861	4	0.0	0	75.0	0.2	8.003096	6.404954e+01
20	16.0	17.0	18.0	649.0	22.0	16.744222	17	15.0	17	7.0	0.7	1.218138	1.483859e+00
21	2.0	2.0	4.0	649.0	4.0	2.514638	2	0.0	2	4.0	0.4	1.134552	1.287208e+00
22	1.0	2.0	3.0	649.0	4.0	2.306626	2	0.0	2	4.0	0.4	1.099931	1.209848e+00
23	1.0	1.0	2.0	649.0	4.0	1.568567	1	1.0	1	3.0	0.2	0.748660	5.604919e-01
24	1.0	2.0	2.0	649.0	4.0	1.930663	2	1.0	2	3.0	0.2	0.829510	6.880861e-01
25	0.0	0.0	0.0	649.0	3.0	0.221880	0	0.0	0	3.0	0.2	0.593235	3.519279e-01
26	4.0	4.0	5.0	649.0	5.0	3.930663	4	1.0	4	4.0	0.2	0.955717	9.133948e-01
27	3.0	3.0	4.0	649.0	5.0	3.180277	3	1.0	3	4.0	0.2	1.051093	1.104796e+00
28	2.0	3.0	4.0	649.0	5.0	3.184900	3	1.0	3	4.0	0.2	1.175766	1.382426e+00
29	1.0	1.0	2.0	649.0	5.0	1.502311	1	1.0	1	4.0	0.2	0.924834	8.553187e-01
30	1.0	2.0	3.0	649.0	5.0	2.280431	2	1.0	1	4.0	0.2	1.284380	1.649632e+00
31	2.0	4.0	5.0	649.0	5.0	3.536210	4	1.0	5	4.0	0.2	1.446259	2.091665e+00
32	0.0	2.0	6.0	649.0	32.0	3.659476	2	0.0	0	32.0	0.2	4.640759	2.153664e+01

Εικόνα 4.1.10 Προσθήκη νέων διαφορετικών πληροφοριών

Εδώ βλέπουμε ότι σε σχέση με το μάθημα των μαθηματικών, στο μάθημα της πορτογαλικής γλώσσας είχαμε 649 εγγραφές αρχικά.

Παρόλο που το σύνολο των δεδομένων αποτελείται μόνο από 32 εγγραφές χωρίς να έχουμε συμπεριλάβει ακόμα τις κατηγορικές μεταβλητές ούτε έχουμε δει τις μεθόδους διαχείρισης τέτοιου είδους δεδομένων ας προσπαθήσουμε να δημιουργήσουμε ένα μοντέλο γραμμικής οπισθοδρόμησης (Linear Regression) ώστε να εξηγήσουμε τον κώδικα και τα αναλύσουμε τα αποτελέσματα.

4.2 Ανάπτυξη μοντέλου οπισθοδρόμησης

4.2.1 Εκτέλεση αλγορίθμου για αριθμητικές μεταβλητές

Import the relevant libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
from sklearn.linear_model import LinearRegression
```

Εικόνα 4.2.1.1 Εισαγωγή βιβλιοθηκών για την εκτέλεση του αλγορίθμου

Η τελευταία κατά σειρά εισαγωγή βιβλιοθήκης sklearn όπως φαίνεται είναι το μοντέλο της γραμμικής οπισθοδρόμησης (Linear Regression).

Στη συνέχεια φορτώνουμε το σύνολο δεδομένων που έχουμε ήδη επεξεργαστεί με τις 32 εγγραφές:

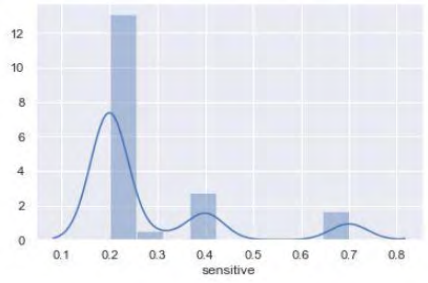
```
# Load the data from a .csv in the same folder
raw_data = pd.read_csv('numerical.csv')

# Let's explore the top 5 rows of the df
raw_data.head()
```

Εικόνα 4.2.1.2 Φόρτωση αριθμητικών μετρικών για τον αλγόριθμο

Το πρώτο βήμα στην εφαρμογή του αλγορίθμου είναι η οπτική απεικόνιση με γραφήματα της ζητούμενης – εξαρτημένης μεταβλητής. Ένα καλό γράφημα είναι αυτό της συνάρτησης πυκνότητας πιθανότητας (Probability Density Function) . Η PDF θα μας βοηθήσει να κατανοήσουμε πως έχει κατανεμηθεί η επιθυμητή μεταβλητή, να εντοπίσουμε αστοχείς (outliers) και ανωμαλίες. Θα δείξουμε το γράφημα PDF για τη ζητούμενη μεταβλητή y, στην περίπτωσή μας είναι η ‘sensitive’ στήλη:

```
# A great step in the data exploration is to display the probability distribution function (PDF) of a variable
# The PDF will show us how that variable is distributed
# This makes it very easy to spot anomalies, such as outliers
# The PDF is often the basis on which we decide whether we want to transform a feature
sns.distplot(raw_data['sensitive'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x157b1e39508>
```



Εικόνα 4.2.1.3 Γράφημα πυκνότητας πιθανότητας για την εξαρτημένη μεταβλητή

Στη συνέχεια προχωράμε στην διευκρίνηση του μοντέλου μας, όπου ανεξάρτητες μεταβλητές είναι και αυτές που θεωρούνται είσοδος (inputs) ενώ ώς έξοδο (target) θα είναι η εξαρτημένη – ζητούμενη μεταβλητή y.

Linear Regression Model

```
#declare inputs and targets
# The target(s) (dependent variable) is 'sensitive'
targets = raw_data['sensitive']

# The inputs are everything BUT the dependent variable, so we can simply drop it
inputs = raw_data.drop(['sensitive'],axis=1)
```

Εικόνα 4.2.1.4 Διαχωρισμός εξαρτημένων και ανεξάρτητων μεταβλητών ως targets και inputs αντίστοιχα

`

Το επόμενο βήμα είναι ο διαχωρισμός των δεδομένων που έχουμε στην κατοχή μας για την εκπαίδευση του συστήματος σε 2 κατηγορίες:

Α) δεδομένα που θα χρησιμοποιηθούν για να εκπαιδεύσουν το σύστημα (x_train και y_train αντίστοιχα για τα inputs και το target). Συνήθως χρησιμοποιείται το 80% του συνόλου.

Β) δεδομένα που θα χρησιμοποιηθούν για να εξεταστεί αν το σύστημα εκπαιδεύτηκε σωστά (x_test και y_test αντίστοιχα). Συνήθως χρησιμοποιείται το 20% του συνόλου.

Για να προχωρήσουμε πρέπει να χρησιμοποιήσουμε τη συνάρτηση της γραμμικής οπισθοδρόμησης που έχουμε εισάγει απ' τη βιβλιοθήκη sklearn και να δημιουργήσουμε ένα νέο αντικείμενο της συνάρτησης αυτής και να εισάγουμε τα δεδομένα για την εκπαίδευση του συστήματος δηλαδή τις εισόδους και την επιθυμητή έξοδο, ως μεταβλητές της συνάρτησης:

```
# Create a linear regression object
reg = LinearRegression()
# Fit the regression with the scaled TRAIN inputs and targets
reg.fit(x_train,y_train)
```

Εικόνα 4.2.1.5 Δημιουργία αντικειμένου γραμμικής οπισθοδρόμησης και τοποθέτηση του training set στον αλγόριθμο

Σε αυτό το σημείο θα αποθηκεύσουμε το αποτέλεσμα της εκπαίδευσης, σε μια μεταβλητή – πίνακα:

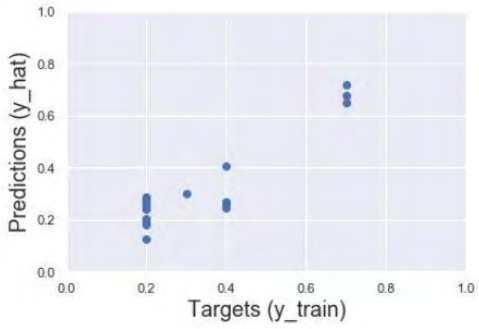
```
# Let's check the outputs of the regression
# I'll store them in y_hat as this is the 'theoretical' name of the predictions
y_hat = reg.predict(x_train)
```

Εικόνα 4.2.1.6 Αποθήκευση αποτελεσμάτων σε ένα πίνακα

4.2.2 Αποτελέσματα και συμπεράσματα εκτέλεσης αλγορίθμου

Τώρα για να δούμε αν τα αποτελέσματα των προβλέψεων ήταν τα επιθυμητά, μπορούμε να χρησιμοποιήσουμε ένα διάγραμμα που στην Y στήλη θα έχει τα αποτελέσματα της εκπαίδευσης y_hat ενώ στη στήλη X θα περιλαμβάνει τις επιθυμητές τιμές της εξαρτημένης μεταβλητής που ορίσαμε προηγουμένως ως y_train. Ένας γενικός κανόνας είναι πως αν έχουμε τις περισσότερες καταγραφές – σημεία (points) στην ευθεία γραμμή των 45^{o} μοιρών, τότε έχει γίνει σωστή πρόβλεψη με μια πρώτη ματιά. Ακολουθεί το διάγραμμα:

```
# The simplest way to compare the targets (y_train) and the predictions (y_hat) is to plot them on a scatter plot
# The closer the points to the 45-degree line, the better the prediction
plt.scatter(y_train, y_hat)
# Let's also name the axes
plt.xlabel('Targets (y_train)',size=18)
plt.ylabel('Predictions (y_hat)',size=18)
# Sometimes the plot will have different scales of the x-axis and the y-axis
# This is an issue as we won't be able to interpret the '45-degree line'
# We want the x-axis and the y-axis to be the same
plt.xlim(0,1)
plt.ylim(0,1)
plt.show()
```

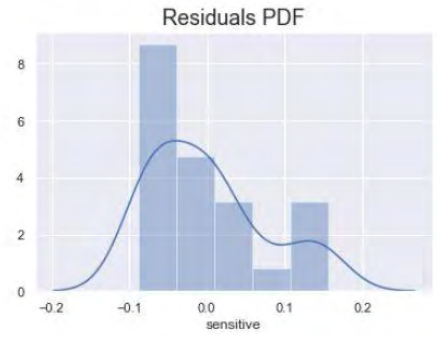


Εικόνα 4.2.2.1 Γράφημα αποτελεσμάτων

Μια ακόμη απεικόνιση μπορεί να είναι το γράφημα της διαφοράς ή αλλιώς Residual PDF. Αυτό που θέλουμε να διαπιστώσουμε είναι αν έχουμε κανονική κατανομή:

```
In [20]: # Another useful check of our model is a residual plot
         # We can plot the PDF of the residuals and check for anomalies
         sns.distplot(y_train - y_hat)

         # Include a title
         plt.title("Residuals PDF", size=18)

         # In the best case scenario this plot should be normally distributed
         # In our case we notice that there are many negative residuals (far away from the mean)
         # Given the definition of the residuals (y_train - y_hat), negative values imply
         # that y_hat (predictions) are much higher than y_train (the targets)
         # This is food for thought to improve our model
```

```
Out[20]: Text(0.5, 1.0, 'Residuals PDF')
```



Εικόνα 4.2.2.2 Γράφημα PDF διαφοράς

Εδώ βλέπουμε ότι έχουμε και αρνητικές τιμές στο γράφημα που σημαίνει ότι οι τιμές που πρόβλεψε το μοντέλο γραμμικής οπισθοδρόμησης (y_hat) είναι μεγαλύτερες από τις επιθυμητές (y_train).

Στη συνέχεια θα δούμε πως μπορούμε να συσχετίσουμε τις εισόδους (Features) με τα βάρη τους (Weights – Coefficients):

Αρχικά, υπολογίζουμε τα coefficients:

```
# Obtain the weights (coefficients) of the regression
reg.coef_

# Note that they are barely interpretable if at all
```

```
array([-5.00806743e-01, -1.63470471e+00, -3.55430158e+00, -1.45545828e-02,
        8.31919408e-01,  1.62967343e+01, -1.63470471e+00,  2.09433969e+01,
        5.08810834e-01, -7.55035997e-01, -1.56841279e+01,  2.72741254e+01])
```

Εικόνα 4.2.2.3 Βάρη των ανεξάρτητων μεταβλητών

Στη συνέχεια, δημιουργούμε ένα πίνακα για να συσχετίσουμε τα features με τα weights τους:

```
# Create a regression summary where we can compare them with one-another
reg_summary = pd.DataFrame(inputs.columns.values, columns=['Features'])
reg_summary['Weights'] = reg.coef_
reg_summary
```

	Features	Weights
0	25%	-0.500807
1	50%	-1.634705
2	75%	-3.554302
3	count	-0.014555
4	max	0.831919
5	mean	16.296734
6	median	-1.634705
7	min	20.943397
8	mode	0.508811
9	range	-0.755036
10	std	-15.684128
11	variance	27.274125

Εικόνα 4.2.2.4 Τοποθέτηση σε πίνακα

Αν υποθέσουμε ότι έχουμε εκπαιδεύσει σωστά το σύστημα μας τώρα πρέπει να κάνουμε τεστ για να δούμε αν τα αποτελέσματα της εκπαίδευσης ήταν θετικά. Προηγουμένως είχαμε χωρίσει το συνολικό μέγεθος των δεδομένων ακριβώς για αυτό το σκοπό. Ενώ το 80% το εκπαιδεύσαμε χρησιμοποιώντας γραμμική οπισθοδρόμηση το 20% θα το χρησιμοποιήσουμε για τη δοκιμή του συστήματος.

Τέλος, θα δούμε τις πραγματικές τιμές της ζητούμενης τιμής y, δηλαδή κατά πόσο μια πληροφορία είναι ευαίσθητη σύμφωνα με τη θεωρία μας:

```
# Finally, let's manually check these predictions
# To obtain the actual sensitive values:
df_pf = pd.DataFrame(y_hat_test, columns=['Prediction'])
df_pf.head()
```

	Prediction
0	0.28
1	0.25
2	0.28
3	0.27
4	0.26

Εικόνα 4.2.2.5 Τιμές που πρόβλεψε ο αλγόριθμος

Συγκρίνουμε την πρόβλεψη μας με την αρχική (πραγματική τιμή) της ζητούμενης μεταβλητής y:

```
# We can also include the test targets in that data frame (so we can manually compare them)
df_pf['Target'] = (y_test)
df_pf

# Note that we have a lot of missing values
# There is no reason to have ANY missing values, though
# This suggests that something is wrong with the data frame / indexing
```

	Prediction	Target
0	0.28	0.20
1	0.25	0.20
2	0.28	0.20
3	0.27	0.20
4	0.26	0.20
5	0.27	0.20
6	0.30	0.20

Εικόνα 4.2.2.6 Σύγκριση τιμών πρόβλεψης με πραγματικών τιμών

Μπορεί να φαίνεται πως η διαφορά δεν είναι μεγάλη και ότι έχει λειτουργήσει ο αλγόριθμος αλλά ας δούμε μια πιο στατιστική προσέγγιση. Υπολογίζουμε τη διαφορά μεταξύ Target και Prediction και βάζουμε και μια επιπλέον μεταβλητή την ποσοστιαία διαφορά (%) για να έχουμε μια ολοκληρωμένη εικόνα των αποτελεσμάτων:

```
# Additionally, we can calculate the difference between the targets and the predictions
# Note that this is actually the residual (we already plotted the residuals)
df_pf['Residual'] = df_pf['Target'] - df_pf['Prediction']

# Since OLS is basically an algorithm which minimizes the total sum of squared errors (residuals),
# this comparison makes a lot of sense
```

Παρακάτω βλέπουμε πως η ποσοστιαία διαφορά είναι υψηλή:

```
# Finally, it makes sense to see how far off we are from the result percentage-wise
# Here, we take the absolute difference in %, so we can easily order the data frame
df_pf['Difference%'] = np.absolute(df_pf['Residual']/df_pf['Target']*100)
df_pf
```

	Prediction	Target	Residual	Difference%
0	0.28	0.20	-0.08	42.03
1	0.25	0.20	-0.05	26.19
2	0.28	0.20	-0.08	39.81
3	0.27	0.20	-0.07	36.42
4	0.26	0.20	-0.06	31.94
5	0.27	0.20	-0.07	33.77
6	0.30	0.20	-0.10	47.93

Εικόνα 4.2.2.7 Πίνακας που αποδίδει την ποσοστιαία διαφορά της πρόβλεψης με πραγματικής τιμής

Για να μπορεί να εκπαιδευτεί αποτελεσματικά το μοντέλο μας πρέπει να μαζέψουμε πληροφορίες – δεδομένα από όσες περισσότερες διαφορετικές πηγές είναι δυνατό. Με αυτό το σκεπτικό προσθέτουμε επιπλέον αριθμητά δεδομένα από πηγή δεδομένων με σκοπό αν το ετήσιο εισόδημα ξεπερνά τις 50000 ή όχι. Πρέπει να αναφέρουμε πως δεν στεκόμαστε στο γεγονός ότι οι πηγές δεδομένων δεν έχουν κοινή λογική διότι εμείς βασιζόμαστε στα μοναδικά ιδιαιτέρα χαρακτηριστικά και στις ιδιότητες κάθε στήλης ξεχωριστά για να συμπεράνουμε αν μια πληροφορία είναι ευαίσθητη ή όχι. Προσθέτουμε επιπλέον μια συλλογή δεδομένων που αφορά την ποιότητα κρασιού. Πλέον στο κατάλληλα διαμορφωμένο αριθμητικό σύνολο δεδομένων μας έχουμε 51 εγγραφές. Επιπλέον προσθέτουμε στο διαμορφωμένο σύνολο μας και άλλες πληροφορίες από πηγή που αφορά την αδικαιολόγητη απουσία εργαζομένων. Πλέον συνολικά το κατάλληλα διαμορφωμένο σύνολο δεδομένων μας περιλαμβάνει 70 αριθμητικές εγγραφές με στήλες που προσδιορίζουν τα γνωρίσματα κάθε αρχικής στήλης.

Εκτελώντας τον ίδιο κώδικα για το μοντέλο της γραμμικής οπισθοδρόμησης όπως έχουμε δείξει παραπάνω αλλά με τις επιπρόσθετες πληροφορίες τα αποτελέσματα έχουν ως εξής:

	Prediction	Target	Residual	Difference%
4	0.29	0.25	-0.04	15.06
12	0.29	0.25	-0.04	16.32
11	0.31	0.40	0.09	21.57
8	0.30	0.40	0.10	25.31
2	0.29	0.40	0.11	26.37
10	0.29	0.40	0.11	26.65
0	0.29	0.40	0.11	27.02
6	0.29	0.40	0.11	28.04
13	0.27	0.20	-0.07	35.21
5	0.36	0.25	-0.11	45.14
9	0.30	0.20	-0.10	52.16
3	0.17	0.40	0.23	56.68
7	0.30	0.70	0.40	57.60
1	0.39	0.20	-0.19	96.53

Εικόνα 4.2.2.8 Προσθήκη συγκρίσεων

Εδώ διαπιστώνουμε ότι έχουμε παραπάνω δοκιμές εφόσον έχουμε και περισσότερες εγγραφές αλλά επίσης ότι σε μερικές δοκιμές η διαφορά από το στόχο κυμαίνεται από 15% έως 96.5 %.

Σε αυτό το σημείο θα δούμε τα αποτελέσματα των εξισώσεων για την αξιολόγηση του μοντέλου γραμμικής οπισθοδρόμησης. Με τη βοήθεια της sklearn βιβλιοθήκης έχουμε τα εξής αποτελέσματα από τη συνάρτηση metrics:

```
: metrics.mean_absolute_error(y_test,y_hat_test)##MAE
: 0.09491898953815685

: metrics.mean_squared_error(y_test,y_hat_test) ##MSE
: 0.01935971607095536

: np.sqrt(metrics.mean_squared_error(y_test,y_hat_test)) ##SMSE
: 0.13913919674540082
```

Εικόνα 4.2.2.9 Αποτελέσματα μετρικών αξιολόγησης μοντέλου

4.3 Προσθήκη και προεπεξεργασία κατηγορικών μεταβλητών

Μέχρι στιγμής έχουμε επεξεργαστεί τις αριθμητικές μεταβλητές αποκτώντας χρήσιμα στατιστικά στοιχεία που στην προσέγγισή μας θεωρούνται inputs - metrics στον αλγόριθμο οπισθοδρόμησης. Για να κατορθώσουμε κάτι αντίστοιχο με τα κατηγορικά δεδομένα, μη αριθμητικά δηλαδή, μπορούμε να εκμεταλλευτούμε μερικές ιδιότητες που παρέχει η Python, δηλαδή ενσωματωμένες λειτουργίες σε δεδομένα μορφής Data Frame ώστε να μπορέσουμε να βγάλουμε μετρικές σημαντικές για το εγχείρημα μας.

Αρχικά, όταν μιλάμε για κατηγορικά δεδομένα η Python τα αναγνωρίζει ως objects – αντικείμενα. Μέσω της συνάρτησης str() μπορούμε να διακρίνουμε αν αυτό το αντικείμενο έχει μόνο αριθμούς, μόνο κεφαλαία, μόνο πεζά, να βρούμε το μέσο όρο από τους χαρακτήρες και φυσικά κατά πόσο η πληροφορία είναι ευαίσθητη με κλίμακα από το 0 ως το 1. Τα αποτελέσματα θα είναι εδώ μια τιμή, 0 ή 1, ανάλογα αν το αποτέλεσμα είναι αληθής η ψευδής. Με αυτό τον τρόπο θα μπορέσουμε να ενσωματώσουμε αυτές τις νέες μετρικές στον αλγόριθμό μας και στη συνέχεια να διακρίνουμε τα νέα αποτελέσματα.

Ξεκινώντας από το πρώτο dataset που επεξεργαστήκαμε και τα αριθμητικά δεδομένα, δηλαδή ένα σύνολο δεδομένων που αφορά πελατολόγιο μιας τράπεζας για την αξιολόγηση και την επιλογή αποδεκτών δανείων ή όχι θα προσπαθήσουμε να υπολογίσουμε τις ζητούμενες μετρικές.

Ως πρώτο βήμα σε κάθε προσπάθεια προεπεξεργασίας δεδομένων, είναι η φόρτωση των κατάλληλων βιβλιοθηκών και του συνόλου:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

Εικόνα 1.1 Εισαγωγή κατάλληλων βιβλιοθηκών

Read the dataset from csv using pandas as DataFrame

```
raw_data = pd.read_csv("bank-full.csv",sep=';')
```

Εϊκόνα 4.3.2 Φόρτωση πρώτου συνόλου δεδομένων σε μορφή .csv

Στη συνέχεια διακρίνουμε τις στήλες που αποτελείται το σύνολο αυτό και να διακρίνουμε τις αριθμητικές τιμές από τις κατηγορικές:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Εικόνα 4.3.3 Εμφάνιση των πρώτων πέντε γραμμών του συνόλου με την συνάρτηση head()

Εδώ μέσω της εντολής head(), εμφανίζονται οι πρώτες πέντε (5) γραμμές με όλες τις στήλες που περιλαμβάνει το σύνολο δεδομένων. Για να δούμε όμως ποια δεδομένα είναι κατηγορικά και ποια όχι, η python διαθέτει την εντολή info() που εκτυπώνει τον τύπο κάθε στήλης χωριστά, που μας βοηθά στον ζητούμενο διαχωρισμό:

```
raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 16 columns):
age          45211 non-null int64
job          45211 non-null object
marital      45211 non-null object
education    45211 non-null object
default      45211 non-null object
balance      45211 non-null int64
housing      45211 non-null object
loan         45211 non-null object
contact      45211 non-null object
day          45211 non-null int64
month        45211 non-null object
duration     45211 non-null int64
campaign     45211 non-null int64
pdays        45211 non-null int64
previous     45211 non-null int64
poutcome     45211 non-null object
dtypes: int64(7), object(9)
```

Εικόνα 4.3.4 Πληροφορίες σχετικά με τον τύπο δεδομένων με τη συνάρτηση info()

Οπότε διακρίνουμε μία μία τις στήλες ως προς τον τύπο των δεδομένων και στο τέλος διακρίνουμε ότι 9 στήλες από τις 16 στο σύνολο έχουν την ένδειξη object. Οπότε ξέρουμε πλέον ποιες πληροφορίες μπορούμε να επεξεργαστούμε ως κατηγορικές μεταβλητές.

Η πρώτη στήλη που διακρίνουμε ως κατηγορική είναι η 'job', δηλαδή η εργασία κάθε πελάτη της συγκεκριμένης τράπεζας. Όταν επεξεργαζόμασταν αριθμητικά δεδομένα είχαμε αναφέρει πως η συνάρτηση describe() εμφανίζει στοιχεία που αφορούν μόνο αριθμητικά στοιχεία. Αν σε αυτή τη συνάρτηση συμπεριλάβουμε την εντολή include=all, τότε θα εμφανίσει στοιχεία και για κατηγορικά δεδομένα και σε πρώτο στάδιο θα εκμεταλλευτούμε αυτό το γεγονός.

Έτσι, δημιουργούμε ένα πίνακα της εντολής αυτής και έχουμε τα εξής αποτελέσματα:

```
Job = [raw_data['job'].describe(include='all')]
```

```
Job
```

```
[count          45211
 unique            12
 top       blue-collar
 freq             9732
 Name: job, dtype: object]
```

Εικόνα 4.3.5 Άντληση γνωρισμάτων κατηγορηματικής μεταβλητής μέσω της συνάρτησης include()

Όπου:

count = σύνολο γραμμών

unique = μοναδικές τιμές της στήλης job, εδώ βλέπουμε ουσιαστικά 12 διαφορετικά επαγγέλματα

top = το επάγγελμα με τη μεγαλύτερη συχνότητα, εδώ συγκεκριμένα το επάγγελμα blue-collar

freq = η τιμή του επαγγέλματος που συναντάμε περισσότερες φορές, δηλαδή το blue-collar με συχνότητα 9732 εγγραφές.

Όμως για να δημιουργήσουμε ένα σύνολο δεδομένων για κατηγορικά δεδομένα δε βοηθάει η μορφή του πίνακα αλλά η μορφή ενός Data Frame. Για αυτό το λόγο μετασχηματίζουμε τον πίνακα σε Data Frame με τη βοήθεια της Pandas:

```
new_job = pd.DataFrame(Job)
```

```
new_job
```

	count	unique	top	freq
job	45211	12	blue-collar	9732

Εικόνα 4.3.6 Μετασχηματισμός πίνακα σε DataFrame με τη συνάρτηση pd.DataFrame()

Εφόσον έχουμε πλέον ένα Data Frame για τη στήλη Job με τις ιδιότητές της, μπορούμε να προσθέτουμε μετρικές για αυτή την κατηγορική μεταβλητή. Είχαμε αναφέρει πως η str έχει συναρτήσεις με λογικό αποτέλεσμα 0 ή 1 ανάλογα αν είναι αληθές η ψευδής το ερώτημά μας. Ας δούμε λοιπόν αν οι τιμές της Job στήλης είναι αλφαβητικές δηλαδή δεν περιέχουν αριθμούς:

```
raw_data['job'].str.isalpha()
```

```
0         True
1         True
2         True
3        False
4         True
        ...
45206     True
45207     True
45208     True
45209    False
45210     True
Name: job, Length: 45211, dtype: bool
```

Εικόνα 4.3.7 Μέσω της συνάρτησης str.isalpha() εξετάζουμε αν η μεταβλητή έχει μόνο χαρακτήρες ή όχι.

Το αποτέλεσμα στις περισσότερες γραμμές – εγγραφές είναι αληθές αλλά βλέπουμε πως σε λίγες γραμμές είναι ψευδής. Άρα το τελικό αποτέλεσμα, αν θέλουμε μια συνολική άποψη για αυτό το ερώτημα είναι : [αληθής και ψευδής = ψευδής]. Έτσι δημιουργούμε μια νέα στήλη με όνομα ‘alphabetic’ και ως τιμή θα έχει το λογικό αποτέλεσμα της παραπάνω πράξης που είναι ίσο με False = Ψευδής:

we can see that not all variables are alphabetic so we can create a new metric named alphabetic = false

```
new_job['alphabetic'] = False
```

```
new_job
```

	count	unique	top	freq	alphabetic
job	45211	12	blue-collar	9732	False

Εικόνα 4.3.8 Εισαγωγή τιμής False στη στήλη 'alphabetic' του DataFrame που είχαμε δημιουργήσει στο προηγούμενο βήμα

Στην ίδια λογική θα δημιουργήσουμε νέες στήλες που θα είναι τελικά οι είσοδοι στον αλγόριθμό μας για τις κατηγορικές – μη αριθμητικές μεταβλητές.

Ως επόμενο metric ελέγχουμε αν η τιμή της κατηγορικής μεταβλητής είναι ίση με το κενό « » :

```
raw_data['job'].str.isspace()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
45206    False
45207    False
45208    False
45209    False
45210    False
Name: job, Length: 45211, dtype: bool
```

```
new_job['isspace'] = False
```

Εικόνα 4.3.9 Ελέγχουμε αν έχουμε κενή συμβολοσειρά και αποθηκεύουμε το αποτέλεσμα στο DataFrame

Ως αποτέλεσμα της λογικής πράξης Ψευδής και Ψευδής = Ψευδής. Άρα η νέα στήλη isspace θα έχει τιμή False σε αυτή την περίπτωση.

Στη συνέχεια ελέγχουμε αν οι τιμές της μεταβλητές είναι αριθμητικές.

```
raw_data['job'].str.isalnum()
```

```
0         True
1         True
2         True
3        False
4         True
         ...
45206     True
45207     True
45208     True
45209    False
45210     True
Name: job, Length: 45211, dtype: bool
```

Εικόνα 4.3.10 Έλεγχος για αριθμητικές τιμές

Πάλι το λογικό αποτέλεσμα είναι False, άρα και η στήλη αντίστοιχα isnum θα έχει τιμή False.

```
new_job['isnum'] = False
```

Εικόνα 4.3.11 Εισαγωγή αποτελέσματος στο DataFrame

Ακόμη ένα χρήσιμο χαρακτηριστικό που μπορεί να θεωρηθεί μετρική στον αλγόριθμό μας είναι αν η συμβολοσειρά μοιάζει με τίτλο, δηλαδή ξεκινά με κεφαλαίο γράμμα και όλα τα υπόλοιπα είναι πεζά:

```
raw_data['job'].str.istitle()
```

```
0         False
1         False
2         False
3         False
4         False
         ...
45206     False
45207     False
45208     False
45209     False
45210     False
Name: job, Length: 45211, dtype: bool
```

```
new_job['title'] = False
```

Εικόνα 4.3.12 Έλεγχος αν η συμβολοσειρά ξεκινά με κεφαλαίο γράμμα

Άλλα χαρακτηριστικά που θα δούμε παρακάτω με όμοιο τρόπο είναι αν τα γράμματα είναι όλα κεφαλαία ή όλα πεζά, αν έχει δεκαδικούς αριθμούς και αν είναι όλα ψηφία:

```
raw_data['job'].str.isdigit()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
45206    False
45207    False
45208    False
45209    False
45210    False
Name: job, Length: 45211, dtype: bool
```

Εικόνα 4.3.13 Έλεγχος αν όλοι οι χαρακτήρες είναι ψηφία

```
raw_data['job'].str.islower()
```

```
0        True
1        True
2        True
3        True
4        True
         ...
45206    True
45207    True
45208    True
45209    True
45210    True
Name: job, Length: 45211, dtype: bool
```

```
new_job['lowercase'] = True
```

Εικόνα 4.3.14 Έλεγχος αν όλα τα γράμματα είναι πεζά

```
raw_data['job'].str.isupper()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
45206    False
45207    False
45208    False
45209    False
45210    False
Name: job, Length: 45211, dtype: bool
```

```
new_job['uppercase'] = False
```

Εικόνα 4.3.15 Έλεγχος αν όλα τα γράμματα είναι κεφαλαία

```
raw_data['job'].str.isdecimal()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
45206    False
45207    False
45208    False
45209    False
45210    False
Name: job, Length: 45211, dtype: bool
```

Εικόνα 4.3.16 Έλεγχος αν έχουμε δεκαδικούς αριθμούς

Οπότε έχουμε διακρίνει αρκετές μετρικές για την εργασία μας. Άλλο ένα σημαντικό χαρακτηριστικό είναι ο μέσος όρος του πλήθους γραμμάτων κάθε μεταβλητής. Αυτό μπορούμε να το υπολογίσουμε και να το χρησιμοποιήσουμε ως είσοδο του αλγορίθμου με τη βοήθεια της βιβλιοθήκης statistics:

```
len_job = raw_data['job'].str.len()
```

```
1 import statistics as st
2 new_job['average length'] = st.mean(len_job)
```

Εικόνα 4.3.17 Υπολογισμός και αποθήκευση μέσο όρο μήκους γραμμάτων

Αρχικά βρίσκουμε και τοποθετούμε το πλήθος των γραμμάτων κάθε τιμής της μεταβλητής σε ένα πίνακα και με τη συνάρτηση mean της βιβλιοθήκης statistics υπολογίζουμε και τοποθετούμε σε μια νέα στήλη του DataFrame που έχουμε δημιουργήσει το average length.

Πλέον το μόνο που απομένει είναι να προσθέσουμε την εξαρτημένη τιμή y – sensitive σε μια νέα στήλη του DataFrame. Η εργασία ενός ατόμου, σε συνδυασμό με άλλα χαρακτηριστικά μπορεί να οδηγήσει στην ταυτοποίησή του οπότε θεωρούμε πως είναι 35% ή 0.35 ευαίσθητη πληροφορία.

```
new_job['sensitive'] = 0.35
```

Εικόνα 4.3.18 Εισαγωγή εξαρτημένης μεταβλητής στο DataFrame

Με την πρώτη επεξεργασία της πρώτης μεταβλητής - στήλης που δεν περιέχει αριθμητικά στοιχεία μπορέσουμε να δημιουργήσουμε νέες μετρικές και εισόδους για τον αλγόριθμο μας.

Τις ίδιες μετρικές θα χρησιμοποιήσουμε για την επεξεργασία κατηγορικών μεταβλητών ώστε να δημιουργήσουμε ένα νέο σύνολο δεδομένων που θα είναι δημιουργημένο μέσα από φυσικά γνωρίσματα κάθε αρχικής στήλης. Αυτό άλλωστε είναι και το σκεπτικό και η προσέγγιση της εργασίας αυτής. Δε θα διακρίνουμε δηλαδή ευαίσθητες πληροφορίες μέσα από τις ετικέτες στηλών (labels), αλλά μέσα από φυσικές πληροφορίες που ανακτάμε μέσα από την προεπεξεργασία αυτών των δεδομένων.

```
labels_str
```

	alphabetic	average length	count	decimal	digit	freq	isnum	isspace	lowercase	sensitive	title	top	unique	uppercase
job	False	9.485546	45211	False	False	9732	False	False	True	0.35	False	blue-collar	12	False
marital	True	6.832275	45211	False	False	27214	False	False	True	0.35	False	married	3	False
education	True	8.320586	45211	False	False	23202	False	False	True	0.35	False	secondary	4	False
month	True	3.000000	45211	False	False	13766	False	False	False	0.20	False	may	12	False

Εικόνα 4.3.19 Εμφάνιση DataFrame

Με μερικές προσθήκες από αυτές τις πληροφορίες θα αποτελείται το DataFrame. Επειδή όμως ο αλγόριθμος της οπισθοδρόμησης εκτελείται μόνο με αριθμητικές εισόδους μπορεί να χρειαστεί να μετατρέψουμε στη συνέχεια τις λογικές τιμές True και False σε 1 και 0 αντίστοιχα.

Έπειτα από μερικές προσθήκες στις κατηγορικές μεταβλητές έχουμε πλέον 16 εγγραφές από τις επεξεργασμένες πληροφορίες που θέλουμε για το μοντέλο της οπισθοδρόμησης:

	alphabetic	average length	count	decimal	digit	freq	isnum	isspace	lowercase	sensitive	title	top	unique	uppercase
0	False	9.485546	45211	False	False	9732	False	False	True	0.35	False	blue-collar	12	False
1	True	6.832275	45211	False	False	27214	False	False	True	0.35	False	married	3	False
2	True	8.320586	45211	False	False	23202	False	False	True	0.35	False	secondary	4	False
3	True	3.000000	45211	False	False	13766	False	False	False	0.20	False	may	12	False
4	False	6.460759	395	False	False	141	False	False	True	0.35	False	other	5	False
5	False	6.136709	395	False	False	217	False	False	False	0.35	False	other	5	False
6	False	6.420253	395	False	False	145	False	False	True	0.20	False	course	4	False
7	False	7.864408	32561	False	False	22696	False	False	False	0.20	False	Private	9	False
8	False	8.433709	32561	False	False	10501	False	False	False	0.30	False	HS-grad	16	False
9	False	14.414054	32561	False	False	14976	False	False	False	0.30	False	Married-civ-spouse	7	False
10	False	12.201898	32561	False	False	4140	False	False	False	0.20	False	Prof-specialty	15	False
11	False	9.119744	32561	False	False	13193	False	False	False	0.30	False	Husband	6	False
12	True	5.538988	32561	False	False	27816	False	False	False	0.30	True	White	5	False
13	True	4.661589	32561	False	False	21790	False	False	False	0.30	True	Male	2	False
14	False	12.293848	32561	False	False	29170	False	False	False	0.30	True	United-States	42	False
15	False	4.759190	32561	False	False	24720	False	False	False	0.30	True	<=50K	2	True

Εικόνα 4.3.20 Εισαγωγή διαφορετικών συνόλων στο DataFrame

Σε προηγούμενα βήματα είχαμε επεξεργαστεί και δημιουργήσει ένα σύνολο από μοναδικά γνωρίσματα που αφορούσε αριθμητικές μεταβλητές. Όμως για να είναι ολοκληρωμένο το μοντέλο μας για τον αλγόριθμο, πρέπει να συνδυάσουμε τα γνωρίσματα και των κατηγορικών και των αριθμητικών μεταβλητών στο ίδιο σύνολο δεδομένων.

Αυτό μπορεί να εκτελεσθεί εύκολα μέσω της εντολής append() απλά στη συνέχεια θα πρέπει να επεξεργαστούμε κατάλληλα της Nan τιμές που έχουν δημιουργηθεί ώστε να μην έχουμε προβλήματα κατά το σχεδιασμό και εκτέλεση του αλγορίθμου:

	25%	50%	75%	alphabetic	average length	count	decimal	digit	freq	isnum	...	min	mode	range	sensitive	std	title	top
0	33.0	39.0	48.0	NaN	NaN	45211.0	NaN	NaN	NaN	NaN	...	18.0	32.0	77.0	0.7	10.618762	NaN	NaN
1	72.0	448.0	1428.0	NaN	NaN	45211.0	NaN	NaN	NaN	NaN	...	-8019.0	0.0	110146.0	0.3	3044.765829	NaN	NaN
2	8.0	16.0	21.0	NaN	NaN	45211.0	NaN	NaN	NaN	NaN	...	1.0	20.0	30.0	0.4	8.322476	NaN	NaN
3	103.0	180.0	319.0	NaN	NaN	45211.0	NaN	NaN	NaN	NaN	...	0.0	124.0	4918.0	0.2	257.527812	NaN	NaN
4	1.0	2.0	3.0	NaN	NaN	45211.0	NaN	NaN	NaN	NaN	...	1.0	1.0	62.0	0.2	3.098021	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11	NaN	NaN	NaN	False	9.119744	32561.0	False	False	13193.0	False	...	NaN	NaN	NaN	0.3	NaN	False	Husband
12	NaN	NaN	NaN	True	5.538988	32561.0	False	False	27816.0	False	...	NaN	NaN	NaN	0.3	NaN	True	White
13	NaN	NaN	NaN	True	4.661589	32561.0	False	False	21790.0	False	...	NaN	NaN	NaN	0.3	NaN	True	Male
14	NaN	NaN	NaN	False	12.293848	32561.0	False	False	29170.0	False	...	NaN	NaN	NaN	0.3	NaN	True	United-States
15	NaN	NaN	NaN	False	4.759190	32561.0	False	False	24720.0	False	...	NaN	NaN	NaN	0.3	NaN	True	<=50K

86 rows × 26 columns

Figure 4.3.21 Προσθήκη κατηγορηματικών και αριθμητικών μετρικών στο ίδιο DataFrame

Πλέον το σύνολο μας αποτελείται από 86 εγγραφές και 26 στήλες, από τις οποίες καλούμαστε να επιλέξουμε ποιες θα είναι features στην γραμμική οπισθοδρόμηση

4.4 Διαχείριση NaN τιμών και εκτέλεση αλγορίθμου

Ας δοκιμάσουμε ξανά την εκτέλεση του αλγορίθμου έχοντας πλέον προσθέσει και μετρικές για τα κατηγορικά δεδομένα. Το πρόβλημα που αναμέναμε να συναντήσουμε είναι ότι οι είσοδος στον αλγόριθμο που αποτελείται από όλες τις στήλες πλην της εξαρτημένης – ζητούμενης μεταβλητής [‘sensitive’] περιέχει NaN τιμές που δεν τις αποδέχεται ο αλγόριθμος της γραμμικής οπισθοδρόμησης:

```
# Create a linear regression object
reg = LinearRegression()
# Fit the regression with the scaled TRAIN inputs and targets
reg.fit(x_train,y_train)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-17-6dfeecd7ee5a> in <module>
      2 reg = LinearRegression()
      3 # Fit the regression with the scaled TRAIN inputs and targets
----> 4 reg.fit(x_train,y_train)

~\Anaconda3\lib\site-packages\sklearn\linear_model\base.py in fit(self, X, y, sample_weight)
    461         n_jobs_ = self.n_jobs
    462         X, y = check_X_y(X, y, accept_sparse=['csr', 'csc', 'coo'],
--> 463                          y_numeric=True, multi_output=True)
    464
    465         if sample_weight is not None and np.atleast_1d(sample_weight).ndim > 1:

~\Anaconda3\lib\site-packages\sklearn\utils\validation.py in check_X_y(X, y, accept_sparse, accept_large_sparse, dtype, order,
copy, force_all_finite, ensure_2d, allow_nd, multi_output, ensure_min_samples, ensure_min_features, y_numeric, warn_on_dtype, e
stimator)
    717                     ensure_min_features=ensure_min_features,
    718                     warn_on_dtype=warn_on_dtype,
--> 719                     estimator=estimator)
    720     if multi_output:
    721         y = check_array(y, 'csr', force_all_finite=True, ensure_2d=False,

~\Anaconda3\lib\site-packages\sklearn\utils\validation.py in check_array(array, accept_sparse, accept_large_sparse, dtype, orde
r, copy, force_all_finite, ensure_2d, allow_nd, ensure_min_samples, ensure_min_features, warn_on_dtype, estimator)
    540         if force_all_finite:
    541             _assert_all_finite(array,
--> 542                                allow_nan=force_all_finite == 'allow-nan')
    543
    544     if ensure_min_samples > 0:

~\Anaconda3\lib\site-packages\sklearn\utils\validation.py in _assert_all_finite(X, allow_nan)
     54                 not allow_nan and not np.isfinite(X).all()):
     55             type_err = 'infinity' if allow_nan else 'NaN, infinity'
---> 56             raise ValueError(msg_err.format(type_err, X.dtype))
     57     # for object dtype data, we only check for NaNs (GH-13254)
     58     elif X.dtype == np.dtype('object') and not allow_nan:

ValueError: Input contains NaN, infinity or a value too large for dtype('float64').
```

Εικόνα 4.4.1 Σφάλμα κατά την εκτέλεση του αλγορίθμου λόγω των NaN τιμών

Οπότε στη συνέχεια πρέπει να βρούμε ένα τρόπο αντιμετώπισης των NaN τιμών ώστε να μην επηρεάζει τον αλγόριθμο και τα αποτελέσματα.

Σύμφωνα με άρθρα και πηγές στο διαδίκτυο σχετικά με την αντιμετώπιση του προβλήματος των τιμών για τον αλγόριθμο της γραμμικής οπισθοδρόμησης υπάρχουν αρκετές εναλλακτικές διαχείρισης του θέματος αλλά οι επικρατέστερες είναι οι εξής:

Μια πολύ απλή λύση είναι η αντικατάσταση όλων των NaN τιμών με 0 αλλά είναι αυτή μια αξιόπιστη επιλογή; Αν αποδεχτούμε αυτήν την τεχνική και εκτελέσουμε τον αλγόριθμο μας θα έχουμε τα εξής αποτελέσματα:

	Prediction	Target	Residual	Difference%
8	0.25	0.25	-0.00	1.51
5	0.33	0.35	0.02	4.49
3	0.31	0.30	-0.01	4.67
12	0.32	0.30	-0.02	5.46
17	0.23	0.20	-0.03	17.08
0	0.24	0.20	-0.04	17.52
7	0.24	0.20	-0.04	19.45
16	0.25	0.20	-0.05	25.67
6	0.31	0.25	-0.06	25.83
4	0.29	0.40	0.11	27.04
15	0.25	0.20	-0.05	27.28
9	0.26	0.20	-0.06	29.59
13	0.26	0.20	-0.06	31.22
2	0.20	0.30	0.10	31.82
1	0.26	0.40	0.14	33.94
14	0.22	0.70	0.48	69.21
11	0.08	0.30	0.22	73.00
10	-1218.86	0.20	1219.06	609529.11

Εικόνα 4.4.2 Αποτελέσματα πρόβλεψης με NaN = 0

4.5 Αποτελέσματα και συμπεράσματα υλοποίησης

Επειδή μέχρι τώρα δεν έχουμε εκτελέσει τον αλγόριθμο συσχετίζοντας αριθμητικές και κατηγορικές τιμές, η μόνη έγκυρη πηγή αξιολόγηση του αλγορίθμου είναι τα αποτελέσματα από τις μετρικές MAE, MSE και RMSE:

```
from sklearn import metrics
```

```
metrics.mean_absolute_error(y_test,y_hat_test)##MAE
```

```
67.80839934314801
```

```
metrics.mean_squared_error(y_test,y_hat_test) ##MSE
```

```
82561.29381501208
```

```
np.sqrt(metrics.mean_squared_error(y_test,y_hat_test)) ##RMSE
```

```
287.3348113525615
```

Εικόνα 4.5.1 Αποτελέσματα αξιολόγησης μοντέλου με NaN = 0

Τα αποτελέσματα των παραπάνω δεν είναι ενθυμητικά καθώς διακρίνουμε μεγάλες διαφορές. Ας αλλάξουμε τη μέθοδο συμπλήρωσης κενών τιμών (NaN) αυτή τη φορά χρησιμοποιώντας τη λογική του μέσου όρου (mean):

```
raw_data.fillna(raw_data.mean(), inplace=True)
```

Εικόνα 4.5.2 Αντικατάσταση NaN τιμών με το μέσο όρο

	Prediction	Target	Residual	Difference%
5	0.33	0.35	0.02	4.54
3	0.31	0.30	-0.01	4.91
8	0.26	0.25	-0.01	5.13
12	0.32	0.30	-0.02	5.30
6	0.30	0.25	-0.05	18.06
0	0.24	0.20	-0.04	18.24
17	0.24	0.20	-0.04	19.34
7	0.24	0.20	-0.04	22.41
16	0.25	0.20	-0.05	26.35
15	0.26	0.20	-0.06	27.71
4	0.28	0.40	0.12	29.29
9	0.26	0.20	-0.06	30.25
13	0.26	0.20	-0.06	30.73
2	0.20	0.30	0.10	32.02
1	0.26	0.40	0.14	34.73
14	0.23	0.70	0.47	66.94
11	0.08	0.30	0.22	74.79
10	1.88	0.20	-1.68	842.42

Εικόνα 4.5.3 Αποτελέσματα πρόβλεψης με NaN = mean()

```
from sklearn import metrics
```

```
metrics.mean_absolute_error(y_test,y_hat_test)##MAE
```

```
0.17691824343676707
```

```
metrics.mean_squared_error(y_test,y_hat_test) ##MSE
```

```
0.17621457840281826
```

```
np.sqrt(metrics.mean_squared_error(y_test,y_hat_test)) ##RMSE
```

```
0.41977920196553126
```

Εικόνα 4.5.4 Αποτελέσματα αξιολόγησης μοντέλου με NaN = mean()

Με αυτό τον τρόπο διαχείρισης κενών τιμών (NaN values) το αποτέλεσμα αλλάζει και διαμορφώνεται σε τεράστιο βαθμό.

Σύμφωνα με τα κριτήρια αξιολόγησης του μοντέλου της γραμμικής οπισθοδρόμησης που έχουμε μελετήσει στην ενότητα 3.5.1 του κεφαλαίου 3, έχουμε υπολογίσει τις τρεις (3) πιο συνήθεις και γνωστές μετρικές αξιολόγησης:

1. **Μέσο απόλυτο σφάλμα (Mean Absolute Error) ~= 0,177**

Αυτό σημαίνει ότι έχουμε υπολογίσει την απόλυτη τιμή της διαφοράς (πρόβλεψη μείον πραγματική τιμή της εξαρτημένης μεταβλητής y-sensitive). Επειδή το MAE δεν υποδηλώνει υποαπόδοση ή υπεραπόδοση του μοντέλου, η τιμή του MAE συμβάλλει αναλογικά στο συνολικό ποσό σφάλματος, πράγμα που σημαίνει ότι όσο μικρότερη είναι η τιμή του MAE τόσο καλύτερη δυνατή πρόβλεψη έχει γίνει. Αντίστοιχα, όσο μεγαλύτερη είναι η τιμή τόσο το μοντέλο μας δεν είναι αντιπροσωπευτικό. Η τιμή 0,177 υποδηλώνει ότι το μοντέλο μας δεν είναι ο τέλειος δείκτης πρόβλεψης, αλλά είναι καλός.

2. **Σφάλμα μέσου τετραγώνου (Mean Squared Error) = 0,176**

Αυτό σημαίνει ότι όσο μεγαλύτερη είναι η τιμή τόσο μεγαλύτερο είναι το σφάλμα. Στην περίπτωσή μας έχουμε μικρή τιμή, άρα και μικρό σφάλμα. Στις περισσότερες περιπτώσεις το MSE χρησιμοποιείται για την επιλογή ενός μοντέλου έναντι ενός άλλου.

3. **Σφάλμα ρίζας μέσου όρου τετραγώνου (Root Mean Squared Error) = 0,42**

Με βάση ένα γενικό κανόνα μπορεί να ειπωθεί ότι οι τιμές μεταξύ 0,2 και 0,5 του RMSE υποδεικνύουν ότι το μοντέλο μπορεί να προβλέψει με σχετική ακρίβεια τα δεδομένα. Το RMSE είναι ένα καλό μέτρο για το πόσο ακριβώς το μοντέλο προβλέπει την αποτελεσματικότητα και είναι το πιο σημαντικό κριτήριο όταν ο κύριος σκοπός του μοντέλου είναι η πρόβλεψη. Έχει επίσης τη σημαντική ιδιότητα να βρίσκεται στις ίδιες μονάδες μέτρησης με τη ζητούμενη μεταβλητή – sensitive.

Μέσα απ’ αυτές τις τρεις (3) μετρικές αξιολόγησης του μοντέλου καταλήγουμε στο συμπέρασμα ότι το μοντέλο της γραμμικής οπισθοδρόμησης που χρησιμοποιήσαμε είναι αποδοτικό για το σύνολο δεδομένων που είχαμε στη διάθεσή μας. Μια σημαντική παρατήρηση είναι ότι αν αυξήσουμε το σύνολο των δεδομένων για το μοντέλο μας μπορεί τα αποτελέσματα να αλλάξουν σε μεγάλο βαθμό οδηγώντας μας σε μια καλύτερη ή χειρότερη πρόβλεψη.

5. Συμπεράσματα

Σε αυτήν την εργασία αναλύσαμε τους κανόνες και τις οδηγίες του ΓΚΠΔ, εξετάσαμε και εντοπίσαμε τα προβλήματα που έχει δημιουργήσει στη βιομηχανία σε φορείς και οργανισμούς όσο αφορά την επεξεργασία και την ανάλυση προσωπικών πληροφοριών με σκοπό την ανάπτυξη και την βελτιστοποίηση των δραστηριοτήτων τους.

Λόγω της ταχύρρυθμης ανάπτυξης της τεχνολογίας ο όγκος πληροφοριών που καλείται κάθε φορέας, ιδιωτικός ή δημόσιος, να επεξεργαστεί και να μετασχηματίσει τις ευαίσθητες προσωπικές πληροφορίες που υπάγονται στο GDPR με τις τεχνικές της ανωνυμοποίησης, ψευδοανωνυμοποίησης ή κρυπτογραφίας, είναι τεράστιος ώστε να περατωθεί χειροκίνητα με αποτέλεσμα να σπαταλείται σημαντικός χρόνος προσωπικού και μεγάλη δαπάνη υπολογιστικής ισχύος.

Εμείς προσεγγίσαμε αυτό το πρόβλημα με τη δημιουργία και υλοποίηση αλγορίθμου εποπτευόμενης μηχανικής μάθησης ώστε να εντοπίζονται αυτόματα οι ευαίσθητες πληροφορίες. Με το σκεπτικό αυτό, αν οποιοδήποτε πληροφοριακό σύστημα ενσωματώσει τον αλγόριθμο που κατασκευάσαμε θα είναι σε θέση να ανιχνεύει τέτοιου είδους πληροφορίες.

Συγκεκριμένα θεωρήσαμε πως για να λειτουργεί πιο αποτελεσματικά ο αλγόριθμος τα δεδομένα εισόδου δεν επαρκεί να είναι κάποιες ετικέτες με πιθανό ευαίσθητο περιεχόμενο αλλά αναλύσαμε τα δεδομένα αυτά βρίσκοντας μοναδικές πληροφορίες που χαρακτηρίζουν τη φύση της πληροφορίας.

Παρόλο που οι πηγές πληροφοριών που απαιτούνται σε μια τέτοια εργασία πρέπει να είναι μεγάλος, καταφέραμε με την συνάθροιση διαφορετικών πληροφοριών να συλλέξουμε επαρκείς πληροφορίες ώστε να υλοποιήσουμε τον αλγόριθμο.

Τα αποτελέσματα της διαδικασίας αυτής μας δείχνουν πως ο αλγόριθμος έχει εκπαιδευτεί σωστά, χρήζει βελτιώσεις για να είναι ακόμα πιο αποδοτικός και να λειτουργεί αποτελεσματικά, αλλά ως γενικό συμπέρασμα στο αρχικό ερώτημα που θέσαμε να απαντήσουμε είναι πως είναι εφικτό μέσα από τεχνικές μηχανικής μάθησης να δοθεί λύση σε ένα σύγχρονο και επίκαιρο πρόβλημα που θα διευκολύνει την ομαλή λειτουργία των επιχειρήσεων και οργανισμών σεβόμενοι την ιδιωτικότητα των φυσικών ατόμων.

Βιβλιογραφία

Πηγές Δεδομένων

Σύνολο δεδομένων απουσιών στην εργασία σε εταιρεία ταχυμεταφορών στη Βραζιλία.

[Andrea Martiniano, Ricardo Pinto Ferreira, and Renato Jose Sassi]

https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Σύνολο δεδομένων απογραφής εισόδων.

[Ronny Kohavi and Barry Becker]

http://archive.ics.uci.edu/ml/datasets/Adult

Σύνολο δεδομένων για τη πρόβλεψη βαθμών στο μάθημα των μαθηματικών στη δευτεροβάθμια εκπαίδευση.

[Paulo Cortez, University of Minho, GuimarÃ£es]

https://archive.ics.uci.edu/ml/datasets/student+performance

Σύνολο δεδομένων άμεσου μάρκετινγκ πορτογαλικού τραπεζικού συστήματος

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Σύνολο δεδομένων που σχετίζεται με δείγματα κόκκινων και λευκών κρασιών

[Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009]

https://archive.ics.uci.edu/ml/datasets/wine+quality

Νομοθεσία ΓΚΠΔ – GDPR

[Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)]

https://eur-lex.europa.eu/legal-content/EL/TXT/?uri=CELEX%3A32016R0679

Μηχανική Μάθηση και Python (Jupyter Notebook)

[Perkel, Jeffrey M. "Why Jupyter is data scientists' computational notebook of choice." Nature, vol. 563, no. 7732, 2018, p. 145+.]

https://go.gale.com/ps/anonymous?id=GALE%7CA573082717&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00280836&p=HRCA&sw=w