ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Σχεδιασμός Συστήματος Συλλογής Δεδομένων, Κατασκευή Προγραμματιστικών Διεπαφών και Προβλεπτικών Μεθόδων για Δεδομένα Καταστροφών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΑΓΓΕΛΟΣ Κ. ΒΑΓΙΑΝΟΣ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Σχεδιασμός Συστήματος Συλλογής Δεδομένων, Κατασκευή Προγραμματιστικών Διεπαφών και Προβλεπτικών Μεθόδων για Δεδομένα Καταστροφών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΑΓΓΕΛΟΣ Κ. ΒΑΓΙΑΝΟΣ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Οκτωβρίου 2022.

Δημήτριος Ασκούνης Καθηγητής Ε.Μ.Π. ΗΜΜΥ Ε.Μ.Π.	Χρυσόστομος Δούκας Καθηγητής Ε.Μ.Π. ΗΜΜΥ Ε.Μ.Π.	Ιωάννης Ψαρράς Καθηγητής Ε.Μ.Π. ΗΜΜΥ Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

Ευάγγελος Κ. Βαγιανός
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

Περίληψη

Η συγκεκριμένη διπλωματική εργασία πραγματεύεται το ζήτημα της αναζήτησης και της επιλογής ετερογενών βάσεων δεδομένων καταστροφών. Οι καταστροφές που αναλύθηκαν χωρίζονται σε δύο κατηγορίες, ανάλογα με την αιτία δημιουργίας τους: γεωφυσικές και σχετιζόμενες με τον καιρό. Αρχικά, έπρεπε να εντοπιστούν βάσεις που περιέχουν πληθώρα δεδομένων των καταστροφών αυτών. Πολλοί κρατικοί φορείς και εταιρείες παρέχουν δωρεάν πρόσβαση σε όλους τους χρήστες για επεξεργασία των βάσεών τους και την εξαγωγή ωφέλιμων συμπερασμάτων. Επιλέχτηκαν εννέα βάσεις και έγινε σύγκριση των περιεχομένων τους και της εγκυρότητας τους Τα κριτήρια αφορούν πρωτίστως την πλήρη τεκμηρίωση και την πληρότητα των βάσεων αυτών. Παρατηρήθηκε ότι πολλές βάσεις, οι οποίες φαινομενικά πληρούσαν τα κριτήρια, απορρίφθηκαν λόγω έλλειψης καταχωρήσεων ή λόγω του ότι περιείχαν παρωχημένα δεδομένα. Οι εννέα βάσεις που επιλέχθηκαν αναλύθηκαν εκτενώς με βάση κάποια συγκεκριμένα γνωρίσματά τους. Στη συνέχεια, έγινε μία επιπλέον διαλογή (ξεσκαρτάρισμα) προκειμένου να επιλεχτούν οι 5 πληρέστερες. Με χρήση διάφορων εργαλείων οι βάσεις αυτές απαλλάχτηκαν από τις περιττές πληροφορίες που περιείχαν και διαμορφώθηκαν βάσει ενός συγκεκριμένου επιθυμητού προτύπου. Στη συνέχεια, παρουσιάστηκαν τα δεδομένα των βάσεων, αφού είχε γίνει η απαραίτητη επεξεργασία. Αναλύθηκαν εκτενώς τα συστήματα που χρησιμοποιήθηκαν και ο τρόπος που εφαρμόστηκαν, όπου χρειάστηκε. Έγινε επιλογή πέντε αλγορίθμων ταξινόμησης μηχανικής μάθησης (ML), οι οποίοι αναλύθηκαν εκτενώς με ορισμούς και παραδείγματα. Σε δύο διαφορετικές περιπτώσεις εφαρμόστηκαν και οι πέντε, προκειμένου να γίνει σύγκριση της αποτελεσματικότητάς τους και να εξαχθούν συμπεράσματα για το χρόνο και την πιθανότητα επιτυχίας του κάθε μοντέλου.

Λέξεις κλειδιά:

Βάσεις δεδομένων καταστροφής, Αλγόριθμοι ταξινόμησης μηχανικής μάθησης, Python, Python Pandas, Βελτιστοποίηση, Jupyter Notebook

Abstract

This thesis deals with the issue of searching and selecting heterogeneous disaster databases. The disasters analyzed, are divided into two minor categories according to the cause of their creation: these categories are geophysical and weather-related. Initially, databases containing a lot of data on these disasters had to be identified. Many government agencies and companies provide free access to all users to process their databases and draw useful conclusions. Nine databases were selected, and their contents and validity were compared. The criteria primarily concern the complete documentation and completeness of these databases. It was observed that many bases which apparently met the criteria were rejected due to missing records or because they contained outdated data. The nine selected bases were extensively analyzed based on their specific features. Then, an additional sorting (discarding) was done to select the five most complete. Using Python Pandas, these bases were freed from the redundant information they contained and configured based on a specific desired pattern. Using JSON, the database data were presented after the necessary processing had been done in Jupyter. The systems were analyzed in detail how they are used. Five machine learning (ML) classification algorithms were chosen and analyzed extensively by using definitions and examples. In the ‘’Disaster_Size’’ column of the Global landslide Catalog base, all five were applied to compare their effectiveness and to draw conclusions about the time and probability of success of each model.

Key words:

Disaster Databases, Python, Python Pandas, Optimization, Machine learning, Classification algorithms, Jupyter, Notebook

Ευχαριστίες

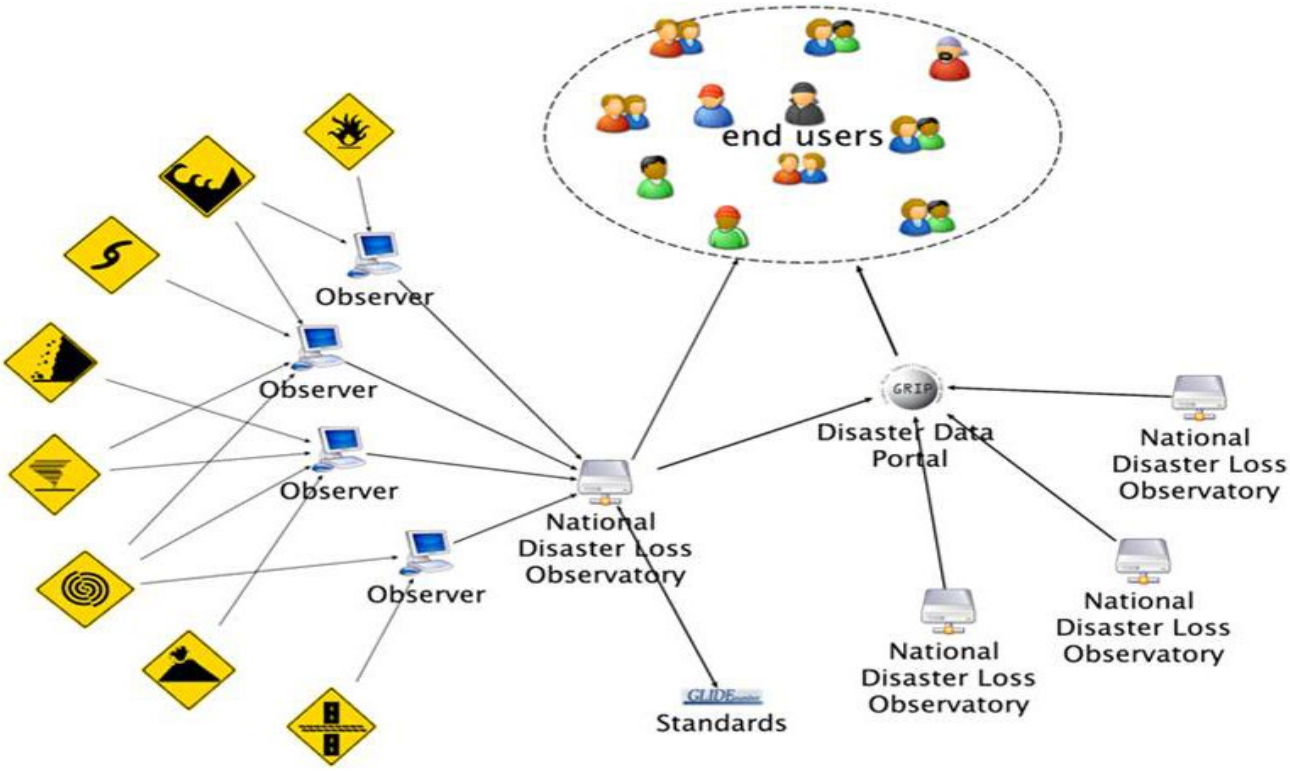
Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο της απόκτησης του πτυχίου του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου.
Θα ήθελα πρωτίστως να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου, καθηγητή Δημήτριο Ασκούνη για την ευκαιρία που μου έδωσε, την εμπιστοσύνη που μου έδειξε αλλά και την καθοδήγησή του .
Επίσης, οφείλω ένα ιδιαίτερο ευχαριστώ στους Υ.Δ. Χριστόδουλο Σαντοριναίο και Ηλιάνα Μάλλιου για τις πολύτιμες συμβουλές τους και την βοήθεια που μου παρείχαν καθ' όλη τη διάρκεια της προετοιμασίας της συγκεκριμένης διπλωματικής εργασίας, μέσα από πολλές δημιουργικές συζητήσεις.
Επιπλέον, στους καθηγητές που είχα ανά τα χρόνια και που με τις γνώσεις που μου μεταλαμπάδευσαν κατέστη εφικτή η διεκπεραίωση και η επιτυχής ολοκλήρωση των σπουδών μου. Τους συμφοιτητές και τους φίλους μου που έκαναν αυτό το μεγάλο ταξίδι, συναρπαστικό και αξέχαστο.
Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια μου για την στήριξη και την αγάπη που μου παρείχε χωρίς ποτέ να σταματήσει να πιστεύει σε μένα.
Σας ευχαριστώ όλους πραγματικά με όλη μου την ψυχή γιατί με βοηθήσατε να βρω τον εαυτό μου και τα θέλω μου με τον πιο δημιουργικό τρόπο.
Ευάγγελος Κ. Βαγιανός
Αθήνα, 25η Οκτωβρίου 2022

Εισαγωγή

Οι φυσικές καταστροφές αποτέλεσαν, από τα πρώτα χρόνια της εμφάνισης του ανθρώπινου είδους ένα μυστήριο. Πολλοί πολιτισμοί απέδωσαν καταστροφές όπως οι σεισμοί, οι κεραυνοί και τα ηφαίστεια σε θεϊκή παρέμβαση ή και τιμωρία. Λόγω αυτού καταλαβαίνουμε τις συνέπειες που αυτές επέφεραν και οδήγησαν μέχρι και στην ολοκληρωτική ισοπέδωση σπουδαίων πολιτισμών όπως η Μινωική Κρήτη και η Πομπηία. Σε μια προσπάθεια του ανθρώπου να επιβιώσει και να αμυνθεί απέναντι σε αυτές που αδυνατούσε να κατανοήσει, οργανώθηκε σε προστατευμένες κοινωνίες και χρησιμοποίησε τα πενιχρά τεχνολογικά μέσα που είχε στη διάθεσή του, προκειμένου να εξηγήσει το άγνωστο. Αυτό ξεκίνησε με την κατανόηση φαινομένων μέσω της απόδοσης σε αυτά φυσικές και απτές ιδιότητες, στερώντας τους την υπεράνθρωπή ερμήνευσή τους. Όσο εξελισσόταν η τεχνολογία με το πέρασμα του χρόνου, τόσο καλύτερα κατάφερνε ο άνθρωπος να προβλέπει και να αντιμετωπίζει τα ακραία αυτά φαινόμενα. Από την παρατήρηση κάποιων μοτίβων, όπως η κίνηση των πουλιών πριν την καταιγίδα στα αρχαία χρόνια, μέχρι τη δημιουργία μοντέλων που μπορούν να προβλέψουν μέσα σε δευτερόλεπτα σεισμικές δονήσεις σε βάθος δεκαετιών, η τεχνολογία έχει σημειώσει άλματα προόδου.

Σκοπός της εργασίας αυτής είναι να συμβάλει στην προσπάθεια, που ξεκίνησε από την αρχαιότητα και έχει φτάσει σήμερα σε σημείο να επιδρά θετικά στο βιοτικό επίπεδο και την ποιότητα ζωής του ανθρώπου σε παγκόσμια κλίμακα. Αυτό επιτυγχάνεται σημαντικά με την αναζήτηση και την τροποποίηση ετερογενών βάσεων δεδομένων καταστροφής. Η πρώτη προσπάθεια καταγραφής και αποθήκευσης βάσεων δεδομένων καταστροφής πραγματοποιήθηκε το 1988 στο Centre for Research on the Epidemiology of Disasters (CRED) από τον Παγκόσμιο Οργανισμό Υγείας (Π.Ο.Υ.) σε συνεργασία με την κυβέρνηση του Βελγίου[1]

Στη μελέτη αυτή, αρχικά επιλέγονται οι βάσεις και εξάγονται δεδομένα τα οποία πληρούν προκαθορισμένα κριτήρια αναζήτησής. Στη συνέχεια δημιουργούνται μοντέλα πρόβλεψης για να αποφευχθούν μελλοντικές καταστροφές με τη χρήση αλγορίθμων μηχανικής μάθησης. Σημειώνεται ότι, οι αλγόριθμοι μηχανικής μάθησης αναπτύχτηκαν από τους Walter Pitts και Warren McCulloch το 1943[2], προκειμένου να χαρτογραφήσουν τις διαδικασίες σκέψης και λήψης αποφάσεων στη μοντελοποίηση των νευρωνικών δικτύων. Συνεπώς, στόχος της διπλωματικής αυτής είναι να συνδυαστούν οι βάσεις καταστροφών με τους αλγορίθμους μηχανικής μάθησης, προκειμένου να εξαχθούν συμπεράσματα για την εύρεση του ιδανικού τρόπου αποθήκευσης και εξαγωγής χρήσιμων δεδομένων για την πρόβλεψη αλλά και την αποτροπή μελλοντικών φυσικών καταστροφών.



Εικόνα 1: Βάση Δεδομένων Καταστροφής [1]

[1] https://www.cdema.org/virtuallibrary/index.php/charim-hbook/use-case-book/9-data-management/9-5-hazard-loss-database

1 Μεθοδολογία αναζήτησης και περιεχόμενα βάσεων καταστροφής

Αρχικά εντοπίστηκαν βάσεις με δεδομένα διαφόρων καταστροφών[3]. Στη συνέχεια έγινε περιορισμός στις πιο πλήρεις και πιο σωστά δομημένες, με βάση συγκεκριμένα κριτήρια επιλογής. Για την εύρεσή τους χρησιμοποιήθηκαν λέξεις-κλειδιά όπως disaster databases, disaster data, managing emergency situations. Έπειτα, έγινε διαχωρισμός με βάση το είδος της κάθε καταστροφής. Συγκεκριμένα βάσεις για: πυρκαγιές, πλημμύρες, σεισμούς, χημικά ατυχήματα, τυφώνες, ολισθήσεις εδάφους, ηφαίστεια, ξηρασίες, ακραίες θερμοκρασίες. Προϋπόθεση ήταν, η κάθε βάση να έχει επαρκή δεδομένα και να είναι πλήρως συμπληρωμένη ως προς τις πληροφορίες που περιείχε. Σημαντικό ρόλο στην επιλογή, και αντίστοιχα στην απόρριψη συγκεκριμένων βάσεων, ήταν η ελεύθερη πρόσβαση στα δεδομένα προς όλους τους χρήστες, χωρίς να υπάρχει κάποιος περιορισμός. Κάθε φυσική καταστροφή που προκύπτει με βάση τα δεδομένα που παρέχονται πρέπει να πληροί κάποια βασικά κριτήρια τα οποία θα αναλυθούν στην συνέχεια.

1.1 Ανάλυση καταστροφών:

Σύμφωνα με τον Π.Ο.Υ[4] (Παγκόσμιος Οργανισμός Υγείας) οι καταστροφές χωρίζονται σε πέντε βασικές κατηγορίες, ανάλογα με την αιτία δημιουργίας τους. Η πρώτη κατηγορία αφορά στις γεωφυσικές καταστροφές και σε αυτήν συμπεριλαμβάνονται όσες δημιουργούνται από απότομες μεταβολές στην μορφολογία του εδάφους. Χαρακτηριστικά παραδείγματα γεωφυσικών καταστροφών αποτελούν οι σεισμοί, οι ηφαιστειακές εκρήξεις και οι κατολισθήσεις. Η δεύτερη κατηγορία είναι οι βιολογικές καταστροφές, όπως οι επιδημίες, με τις οποίες όμως δεν θα ασχοληθούμε καθώς τα δεδομένα τους είναι απρόβλεπτα, άπτονται άλλου επιστημονικού πεδίου, και δεν μπορούν συνεπώς να εξαχθούν συμπεράσματα. Οι άλλες τρείς κατηγορίες, μπορούν να ενταχθούν σε μια γενικότερη ομάδα και είναι οι υδρολογικές, οι μετεωρολογικές και οι κλιματολογικές. Και στις τρεις αυτές υποπεριπτώσεις κοινός λόγος δημιουργίας τους είναι οι απότομες και απρόβλεπτες κλιματικές αλλαγές και οι επιπτώσεις που αυτές δημιουργούνε. Οι κυριότερες και συχνότερες καταστροφές της κατηγορίας αυτής είναι οι πλημμύρες, οι καταιγίδες και οι ξηρασίες. Ακολουθεί μία βαθύτερη ανάλυση των κυριότερων καταστροφών.

Εικόνα 2: Φυσικές Καταστροφές [2]

1.2 Δεδομένα Βάσεων

Στα δεδομένα βάσεων περιλαμβάνεται ο τύπος της καταστροφής, όπως και η ακριβής χρονολογική ημερομηνία του συμβάντος[5]. Επίσης, η ακριβής τοποθεσία που λαμβάνει χώρα σε συγκεκριμένη περιφέρεια, κράτος και πόλη. Για διευκόλυνση στην προσπέλαση των δεδομένων, παρέχεται ο αντίστοιχος αναγνωριστικός κωδικός (ID) για κάθε καταστροφή. Στα δεδομένα συγκαταλέγονται ο τύπος της καταστροφής, ποτέ συνέβη, οι συντεταγμένες της τοποθεσίας, η χώρα και η περιοχή με ακρίβεια χιλιομέτρου, καθώς και η ένταση της καταστροφής. Η ένταση ποικίλλει ανάλογα με το είδος της καταστροφής. Σε περιπτώσεις σεισμικών καταστροφών, χρησιμοποιήθηκε η κλίμακα ρίχτερ ενώ σε κατολισθήσεις, η έκταση που επηρέασε και η αλλαγή που επήλθε στη μορφολογία του εδάφους. Ο διαχωρισμός των φυσικών καταστροφών γίνεται με βάση τις επιπτώσεις προς τον άνθρωπο (αριθμός θανάτων, τραυματιών καθώς και μετατόπιση πληθυσμού). Επίσης καταγράφεται, εκτός από το περιβαλλοντολογικό, και το αντίστοιχο οικονομικό πλήγμα.

[2] https://www.noaa.gov/topic-tags/natural-disasters

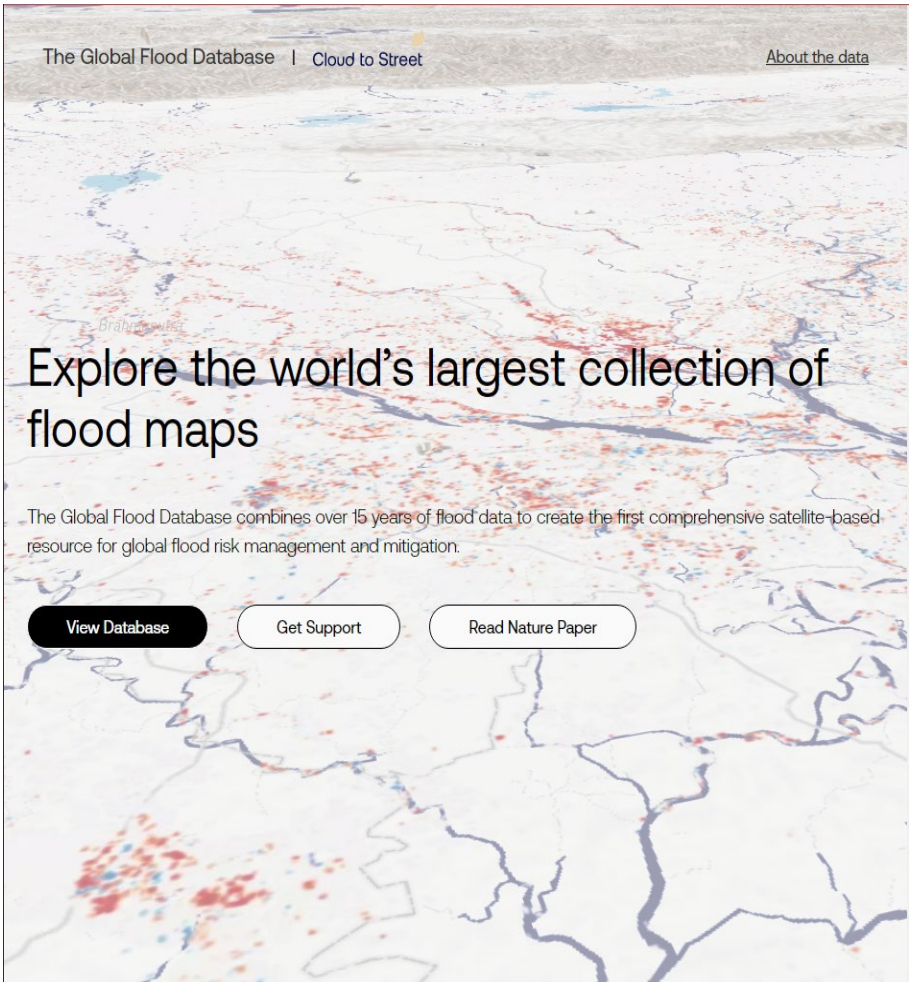
Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

2 Παρουσίαση βάσεων δεδομένων

Οι παρακάτω πίνακες αφορούν στην παρουσίαση των πηγών δεδομένων που έχουν εντοπιστεί κατά την αναζήτηση[6]. Η ανάλυση των πινάκων δεδομένων δίνεται εκτενέστερα στη συνέχεια.

2.1 Global Flood Database

Η πρώτη βάση αφορά πλημμύρες που έχουν συμβεί ανά τον κόσμο. Προσφέρει πλοήγηση ανά χρονολογία ή ημερολογιακά. Υπήρχε η δυνατότητα να γίνει επιλογή της χώρας που επιθυμεί ο χρήστης και στη συνέχεια να εντοπιστεί η αντίστοιχη φυσική καταστροφή. Απεικονιστικά, με τη βοήθεια χάρτη, μπορούσαμε να μελετήσουμε το κέντρο της καταστροφής και την έκταση που επηρεάστηκε από αυτή. Επίσης, ο χρήστης είχε πρόσβαση σε στοιχεία όπως των αριθμό των θυμάτων, τους τραυματίες και τον αντίκτυπο που είχε στον τοπικό πληθυσμό και στα επόμενα έτη.



Εικόνα 3: GFD [3]

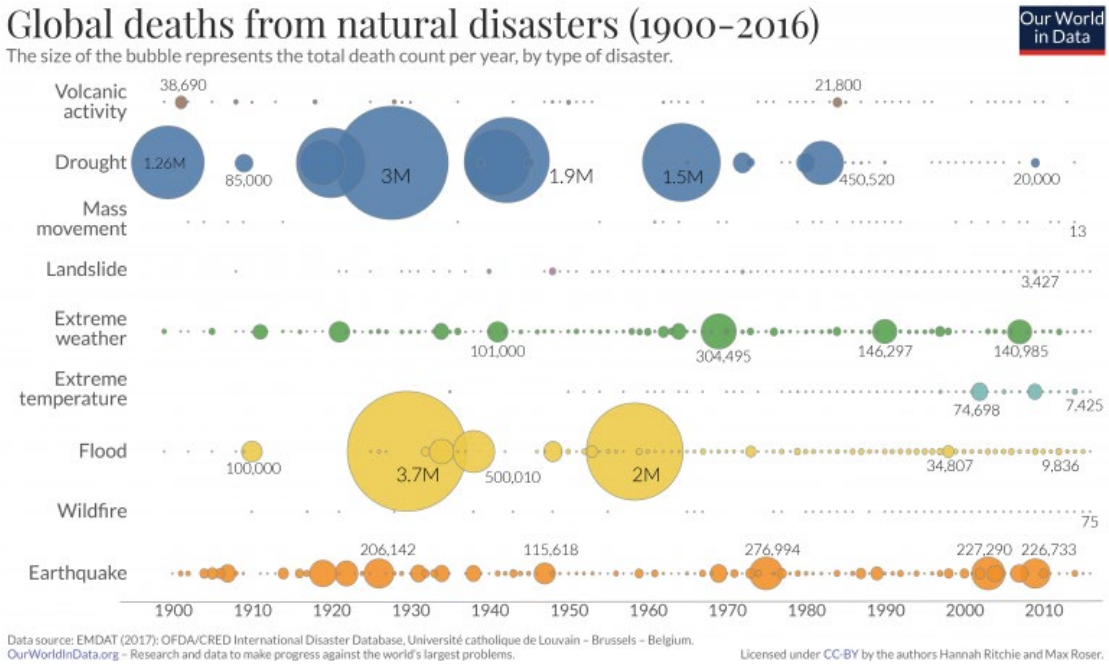
[3] https://global-flood-database.cloudtostreet.ai/

Πίνακας 1: Global Flood Database

Πηγή/Βάση δεδομένων	global-flood database
Περιγραφή	Περιέχει αναφορές πλημμυρών
Είδη καταστροφών που περιλαμβάνονται	Πλημμύρες, τυφώνες.
Κατηγορίες των δεδομένων	Αλφαριθμητικά στοιχεία που αφορούν πλημμύρες και τυφώνες, συμπεριλαμβάνεται ο αριθμός θυμάτων.
Λεπτομέρειες σχετικά με τα δεδομένα	Ημερομηνία που συνέβη η καταστροφή.
	Άνθρωποι που επηρεάστηκαν άμεσα ή έμμεσα.
	Θύματα.
	Εκτοπισμένοι άνθρωποι λόγω του συμβάντος.
	Διάρκεια καταστροφής σε μέρες.
	Αίτια που δημιουργήθηκε (βροχή, καταιγίδα).
Σκοπός	Το global-flood database περιέχει αναφορές για πλημμύρες που έλαβαν χώρα στην Αμερική.
Γλώσσα	Αγγλικά
Δημόσια/Ιδιωτικά δεδομένα	Δημόσια δεδομένα (δεν απαιτείται άδεια για την πρόσβαση σε αυτά ή τη λήψη αυτών).
Type/format των δεδομένων	CSV
Τρόποι εξαγωγής των δεδομένων	Κάθε χρήστης έχει τη δυνατότητα να δει την τοποθεσία, την ημερομηνία και την καταστροφή που θέλει να αναζητήσει. Μπορεί να φιλτράρει τα δεδομένα μέσω ενός φίλτρου αναζήτησης και να ανατρέξει σε έναν χάρτη για πληρέστερη εικόνα. Επίσης έχει διαγραμματική απεικόνιση, που δείχνει την πλήρη έκταση της καταστροφής ανά περιοχή.

2.2 OurWorldInData

Είναι ίσως η πληρέστερη βάση, καθώς παρουσιάζει όλες τις φυσικές καταστροφές, και παράλληλα είναι και εύχρηστη. Επίσης, παρουσιάζονται τα δεδομένα με πολλές μορφές απεικόνισης.
Περιλαμβάνει διαγράμματα, πίνακες και τις πηγές που αντλήθηκαν αυτές οι πληροφορίες. Υπάρχει η επιλογή να "κατεβάσουμε" το αρχείο για προσωπική χρήση.



Εικόνα 4: OURWORLDINDATA [4]

Πίνακας 2: ourworldindata

Πηγή/Βάση δεδομένων	ourworldindata
Περιγραφή	Το ourworldindata περιέχει αναφορές που σχετίζονται με φυσικές καταστροφές ανά χώρα.
Είδη καταστροφών που περιλαμβάνονται	Περιέχει ηφαίστεια, πλημμύρες, σεισμούς, ξηρασίες, ακραίες θερμοκρασίες , ολισθήσεις εδάφους και καταιγίδες.

[4] https://ourworldindata.org/

Κατηγορίες των δεδομένων	Διαλέγοντας την χώρα και την καταστροφή στην αναζήτηση μας, φαίνεται ο αντίκτυπος που είχε στον πληθυσμό και στην οικονομία της. Συγκρίνοντας αριθμούς και βλέπουμε πως άλλαξαν τα δεδομένα ανά έτος.
Λεπτομέρειες σχετικά με τα δεδομένα	Τα δεδομένα μας χωρίζονται σε δύο κύριες κατηγορίες: 1. Θάνατοι, άστεγοι, τραυματίες και γενικότερα πόσοι επηρεάστηκαν από το συμβάν. 2. Οι οικονομικές επιπτώσεις που υπέστη η χώρα που μας ενδιαφέρει.
Σκοπός	Να υπάρξει μια πλήρως ενημερωμένη βάση, εύκολα προσβάσιμη, με δεδομένα που να παρουσιάζονται σε απλή και κατανοητή μορφή αλλά να υπερκαλύπτει τις αναζητήσεις μας.
Γλώσσα	Αγγλικά.
Δημόσια/Ιδιωτικά δεδομένα	Τα δεδομένα είναι δημόσια.
Type/format των δεδομένων	CSV.
Τρόποι εξαγωγής των δεδομένων	Η βάση δεδομένων προσφέρει φίλτρα αναζήτησης και κάθε χρήστης μπορεί είτε να τα "κατεβάσει" σε επεξεργάσιμη μορφή, είτε να τα προωθήσει μέσω κοινωνικών δικτύων.

2.3 BDcatnat

Εικόνα 5: CATNAT [5]

Η βάση 3 παρέχει πληθώρα δεδομένων, αλλά δεν δίνει πλήρη πρόσβαση σε απλούς χρήστες, και απαιτείται οικονομική συνδρομή. Υπάρχει η επιλογή όμως να ՝κατεβάσουμε՝ τα δεδομένα, έστω και με περιορισμένης πρόσβασης χρήση, για τη μελέτη αυτών.

Πίνακας 3: Βάση Catnat

Πηγή/Βάση δεδομένων	catnat.net
Περιγραφή	Το catnat.net μας παρουσιάζει τα αποτελέσματα που θέλουμε με βάση τα κριτήρια που επιλέγουμε. Υπάρχει ταξινόμηση ανά χώρα, ημερομηνία, έκταση του συμβάντος και τύπο καταστροφής.
Είδη καταστροφών που περιλαμβάνονται	Οι καταστροφές χωρίζονται σε πέντε μεγάλες κατηγορίες που αφορούν τον τύπο προέλευσης (κλιματική ,γεωλογική, χωρική, μετεωρολογική, υδρολογική)και σε δεκαοχτώ μικρότερες που αφορούν τον τύπο της (πυρκαγιές, τσουνάμι κ.λπ.)
Κατηγορίες των δεδομένων	Προέλευση καταστροφής.
	Είδος καταστροφής.
	Ήπειρος
	Χώρα
	Έκταση συμβάντος.
Λεπτομέρειες σχετικά με τα δεδομένα	Τη γενικότερη κατηγορία στην οποία εντάσσεται η καταστροφή που αναζητώ.
	Δίνεται το είδος της καταστροφής και έχουμε πληθώρα επιλογών.

[5] https://gdc.unicef.org/resource/bd-catnat-database

	Την ήπειρο στην οποία έλαβε χώρα. Τη χώρα που μας ενδιαφέρει. Το αντίκτυπο που είχε η καταστροφή στο περιβάλλον και τους ανθρώπους που έπληξε.
Σκοπός	Παρέχει δεδομένα για φυσικές καταστροφές που έλαβαν χώρα σε χρονικό διάστημα της επιλογής μας.
Γλώσσα	Αγγλικά, Γαλλικά
Δημόσια/Ιδιωτικά δεδομένα	Η πρόσβαση είναι περιορισμένη για μη συνδρομητές (εμφανίζονται λιγότερα αποτελέσματα), ενώ για τους premium χρήστες υπάρχουν όλα τα δεδομένα.
Type/format των δεδομένων	CSV
Τρόποι εξαγωγής των δεδομένων	Ο χρήστης μπορεί να κατεβάσει τα δεδομένα κάνοντάς τα export to CSV ή EXCEL και στην πορεία να τα κάνει extract. Αυτή τη δυνατότητα την έχουν όλοι οι χρήστες ανεξαρτήτως συνδρομής.

2.4 Copernicus

Η συγκεκριμένη βάση δεν είναι πλήρης και έχει πολύ μικρό αριθμό δεδομένων. Αφορά πυρκαγιές στην Ευρώπη και αντίστοιχα δείχνει, για κάθε χώρα ανά έτος, την έκταση της περιοχής που έπληξε η καταστροφή.

Εικόνα 6: COPERNICUS [6]

[6] https://effis.jrc.ec.europa.eu/apps/effis.statistics/

Πίνακας 4: Copernicus

Πηγή/Βάση δεδομένων	Copernicus
Περιγραφή	Το Europe's eyes on earth παρουσιάζει τα δεδομένα σε μορφή διαγράμματος και χάρτη. Αφορά πυρκαγιές στην Ευρώπη.
Είδη καταστροφών που περιλαμβάνονται	Πυρκαγιές.
Κατηγορίες των δεδομένων	Έτος.
	Έκταση καταστροφής.
	Αριθμός πυρκαγιών.
Λεπτομέρειες σχετικά με τα δεδομένα	Ποια χρονολογία μας ενδιαφέρει.
	Πόσα εκτάρια είναι η καμένη περιοχή
	Πλήθος πυρκαγιών που ξέσπασαν τη συγκεκριμένη χρονιά που μας ενδιαφέρει
Σκοπός	Να κατανοήσουμε καλύτερα το πλήθος και την έκταση των πυρκαγιών στο χρόνο και χώρο που επιθυμούμε
Γλώσσα	Αγγλικά
Δημόσια/Ιδιωτικά δεδομένα	Τα δεδομένα είναι όλα δημόσια και εύκολα προσβάσιμα για οποιονδήποτε χρήστη
Type/format των δεδομένων	JSON, CSV, PDF
Τρόποι εξαγωγής των δεδομένων	Μπορούμε να κάνουμε download τα δεδομένα στις εξής μορφές: Image(PNG, JPG, SVG, PDF), Data(JSON, CSV, PDF), Print

2.5 FireserviceGR

Εικόνα 7: FIRESERVICEGR [7]

Η βάση που παρέχει το fireservice.gr αφορά αποκλειστικά φωτιές που έχουν λάβει χώρα στον Ελλαδικό χώρο. Υπάρχουν πολλές μορφές αρχείων excel και πληροφοριών διάσπαρτα στην ιστοσελίδα που απαρτίζουν την τελική βάση δεδομένων που μελετήθηκε.

Πίνακας 5: FireserviceGR

Πηγή/Βάση δεδομένων	https://www.fireservice.gr/el_GR/synola-dedomenon
Περιγραφή	Το FireserviceGR περιέχει αναφορές πυρκαγιών στον Ελλαδικό χώρο.
Είδη καταστροφών που περιλαμβάνονται	Δασικές και αστικές Πυρκαγιές.
Κατηγορίες των δεδομένων	Αρχείο συμβάντων που επεμβαίνει το Πυροσβεστικο Σωμα.
	Υλικά που έχουν χρησιμοποιηθεί από τους φορείς.
	Εκπαίδευση που παρέχει ο φορέας.
	Κόστη και χρέη που δημιουργήθηκαν λόγω της καταστροφής.
	Επικοινωνία με τους υπεύθυνους του φορέα.
	Λεπτομέρειες σχετικά με τα δεδομένα από το Αρχείο Συμβάντων.
	Πυροσβεστικό Υλικό.
	Τηλεφωνικός Κατάλογος Υπηρεσιών.
	Πιστώσεις.

[7] https://www.fireservice.gr/el/synola-dedomenon

Σκοπός	Το fireservice.gr περιέχει αναφορές πυρκαγιών στην Ελλάδα ανα τα έτη.
Γλώσσα	Ελληνική
Δημόσια/Ιδιωτικά δεδομένα	Σύμφωνα με το Νόμο 4305/2014 επαναπροσδιορίζεται η ανοικτή διάθεση και περαιτέρω χρήση εγγράφων, πληροφοριών και δεδομένων του δημόσιου τομέα, με την προσαρμογή της εθνικής νομοθεσίας στις διατάξεις της Οδηγίας 2013/37/ΕΕ του Ευρωπαϊκού Κοινοβουλίου, για την περαιτέρω ενίσχυση της διαφάνειας.
Type/format των δεδομένων	Αρχεία xl, xls, csv.
Τρόποι εξαγωγής των δεδομένων	Είναι δυνατό για οποιονδήποτε δημόσιο χρήστη να ˙κατεβάσει˙τα δεδομένα, χωρίς να απαιτείται εγγραφή στον φορέα σε μορφή xl, xls, csv.

2.6 Global Internal Displacement Database

Η βάση δεδομένων Global Internal Displacement Database[7] περιέχει δεδομένα για καταστροφές που συνέβησαν μεταξύ 2008-2021 σε παγκόσμια κλίμακα. Συγκεκριμένα, έχει πληροφορίες για 202 χώρες και περιοχές και 342.3 εκατομμύρια μετατοπίσεις πληθυσμού, λόγω γεωφυσικών και καιρικών καταστροφών. Η αναζήτηση των πληροφοριών γίνεται με 4 κριτήρια αναζήτησης: Ήπειρος, χώρα και περιοχή, είδος καταστροφής, χρονικό διάστημα.

Εικόνα 8: GIDD [8]

[8] https://www.internal-displacement.org/

Πίνακας 6: GIDD

Πηγή/Βάση δεδομένων	Global Internal Displacement Database
Περιγραφή	Το Global Internal Displacement Database περιέχει αναφορές συμβάντων ανά τον κόσμο που οδήγησαν σε μετατοπίσεις πληθυσμού στο χρονικό διάστημα 2008-2021
Είδη καταστροφών που περιλαμβάνονται	Γεωφυσικές(σεισμοί, ηφαίστεια) και καταστροφές σχετιζόμενες με τις καιρικές συνθήκες (πλημμύρες, τυφώνες, ακραίες θερμοκρασίες κ.α.)
Κατηγορίες των δεδομένων	Είδος καταστροφής
	Έτος που συνέβη η καταστροφή
	Όνομα καταστροφής
	Ημερομηνία συμβάντος
	Μετατοπίσεις πληθυσμού λόγω του συμβάντος
	Κατηγορία καταστροφής
Λεπτομέρειες σχετικά με τα δεδομένα	Ποια χρονολογία μας ενδιαφέρει (το όνομα της, η χρονολογία και η ημερομηνία που συνέβη)
	Τις αλλαγές στο πλήθος των κατοίκων της περιοχής και τις μετατοπίσεις που προκάλεσε σε αυτό
	Την κατηγορία της καταστροφής (γεωφυσική ή σχετιζόμενη με τον καιρό) και την υποκατηγορία που αφορά τον τύπο της.
Σκοπός	Το Internal-displacement περιέχει δεδομένα για μετατοπίσεις του ευρύτερου πληθυσμού μιας περιοχής και απεικονίζονται τα αίτια που οδήγησαν σε αυτήν. Επίσης, έχει διαγραμματική απεικόνιση αυτών, ανά καταστροφή και ανά έτος.

Γλώσσα Type/format των δεδομένων	Αγγλικά
Δημόσια/Ιδιωτικά δεδομένα	Η διάθεση των δεδομένων της σελίδας είναι ελεύθερη προς όλους
Type/format των δεδομένων	Αρχεία xls, docx
Τρόποι εξαγωγής των δεδομένων	Είναι δυνατό για οποιονδήποτε δημόσιο χρήστη να κατεβάσει τα δεδομένα χωρίς να απαιτείται εγγραφή στον φορέα σε μορφή xls, docx.

2.7 EprecSilence

Η βάση 7 έχει εύκολη πρόσβαση και διαχείριση των δεδομένων που παρέχει. Οι καταστροφές που πραγματεύεται δεν περιορίζονται σε ένα συγκεκριμένο είδος αλλά πολλά διαφορετικά.

Εικόνα 9: EPRECSILENCE [9]

[9] https://www.epcresilience.com/insight/resources/disaster-database

Πίνακας 7: EprecSilence

Πηγή/Βάση δεδομένων	https://www.epcresilience.com/insight/resources/disaster-database.
Περιγραφή	Το https://www.epcresilience.com/insight/resources/disaster-database περιέχει καταστροφές.
Είδη καταστροφών που περιλαμβάνονται	Γεωφυσικές (σεισμοί, ηφαίστεια) και καταστροφές σχετιζόμενες με τις καιρικές συνθήκες (πλημμύρες, τυφώνες, ακραίες θερμοκρασίες κ.α.).
Κατηγορίες των δεδομένων	Έτος που συνέβη η καταστροφή.
	Είδος καταστροφής.
	Χώρα/Πόλη.
	Τύπος καταστροφής.
	Θύματα.
	Τραυματίες.
	Επιπλέον σημειώσεις για την καταστροφή.
	Link(1-4) με πηγές.
	Αναφορές σε βιβλία.
Λεπτομέρειες σχετικά με τα δεδομένα	Ποιά χρονολογία μας ενδιαφέρει.
	Το είδος και ο τύπος της καταστροφής.
	Η τοποθεσία(χώρα και πόλη) που συνέβη.
	Οι πληγέντες: θύματα και τραυματίες (σε απόλυτο αριθμό ή ποσοστιαία).
	Σημειώσεις που προσφέρουν επιπλέον πληροφορίες για την καταστροφή. Παρουσιάζονται τα άμεσα ή έμμεσα αίτια δημιουργίας της και οι ευρύτερες επιπτώσεις που σχετίζονται με την καταστροφή.
	Παρέχεται η δυνατότητα να ανατρέξει ο χρήστης σε πηγές (links ή βιβλιογραφικές αναφορές).
Σκοπός	Το epcresilience.com δίνει τη δυνατότητα να κατεβούν σε μορφή xls, disaster data για ελεύθερη χρήση από όποιον επιθυμεί, με σκοπό την εκτενέστερη διερεύνησή τους και την εξαγωγή χρήσιμων συμπερασμάτων.
Γλώσσα	Αγγλικά.

Δημόσια/Ιδιωτικά δεδομένα	Η διάθεση των δεδομένων της σελίδας είναι ελεύθερη προς όλους .
Type/format των δεδομένων	Αρχεία xls.
Τρόποι εξαγωγής των δεδομένων	Είναι δυνατό για οποιονδήποτε δημόσιο χρήστη να κατεβάσει τα δεδομένα χωρίς να απαιτείται εγγραφή στον φορέα σε μορφή xls, και η βάση ανανεώνεται συνεχώς.

2.8 Global Landslide Catalog

Η βάση[8] έχει 11.033 καταχωρήσεις και 31 στήλες, πληθώρα δεδομένων τα οποία οποιοσδήποτε χρήστης μπορεί να αποκτήσει και να τα τροποποιήσει, ανάλογα με το όφελος που αποσκοπεί να αποκομίσει από αυτά.

10

Εικόνα 10: GLC

[10] https://data.nasa.gov/

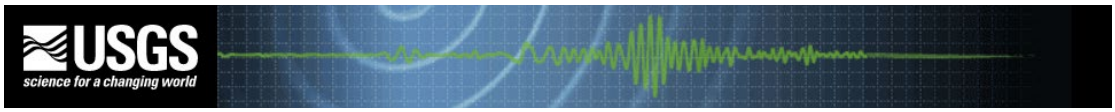
Πίνακας 8: GLC

Πηγή/Βάση δεδομένων	https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v/data
Περιγραφή	Το data.nasa.gov περιέχει δεδομένα που παρέχει η NASA και αφορά ποίκιλλα είδη καταστροφών σε παγκόσμια κλίμακα.
Είδη καταστροφών που περιλαμβάνονται	Κάθε καταστροφή έχει ένα μοναδικό αναγνωριστικό (ID). Περιορίζεται σε φυσικές καταστροφές όπου έχουν αρνητικό αντίκτυπο είτε σε άψυχο είτε σε έμψυχο δυναμικό.
Κατηγορίες των δεδομένων	Η καταστροφή έχει ένα μοναδικό αναγνωριστικό (ID) και μία πηγή.
	Ημερομηνία ατυχήματος, καταχώρησης στην βάση και τελευταίας τροποποίησης.
	Τοποθεσία συμβάντος (Χώρα, Πόλη, Περιφέρεια).
	Θύματα.
	Τραυματίες.
	Εμβέλεια σε χιλιόμετρα.
	Αλλαγές που επήλθαν στη μορφολογία του εδάφους.
Λεπτομέρειες σχετικά με τα δεδομένα	Το αναγνωριστικό αντιστοιχεί σε μία συγκεκριμένη καταστροφή ως μοναδικό γνώρισμα.
	Δίνονται οι συντεταγμένες της τοποθεσίας, η χώρα και το αναγνωριστικό της, η περιοχή και η περιφέρεια αναζήτησης με ακρίβεια χιλιομέτρου.
	Οι πηγές που παρέχονται δίνουν παραπάνω αξιοπιστία στα δεδομένα.
	Υπάρχει μία συνοπτική/περιεκτική και μία πιο αναλυτική περιγραφή της καταστροφής.
	Οι ανθρώπινες απώλειες και οι τραυματίες.

	Οι αλλαγές που επήλθαν στην μορφολογία του εδάφους (βύθιση, κλίση, μείωση μεγέθους κλπ.).
Σκοπός	Το data.nasa.gov παρέχει αξιόπιστα δεδομένα προς επεξεργασία και προς εξαγωγή συμπερασμάτων. Έχει έγκυρες πηγές, άρα αξιοπιστία, και πληθώρα πληροφοριών για τα δεδομένα που προσφέρει.
Γλώσσα	Αγγλικά.
Δημόσια/Ιδιωτικά δεδομένα	Η διάθεση των δεδομένων της σελίδας είναι ελεύθερη προς όλους.
Type/format των δεδομένων	Αρχεία CSV, JSON, xl, XML.
Τρόποι εξαγωγής των δεδομένων	Είναι δυνατό για οποιονδήποτε δημόσιο χρήστη να κατεβάσει τα δεδομένα χωρίς να απαιτείται εγγραφή στον φορέα σε μορφή CSV, JSON, xl, XML, και η βάση ανανεώνεται συνεχώς.

2.9 United States Geological Survey

Η συγκεκριμένη βάση περιείχε 9721 καταχωρήσεις σεισμών[9]. Πολλά δεδομένα κρίθηκαν περιττά και έγινε εφαρμογή φίλτρου επιλογής όσων βάσεων είχαν πάνω από 4 βαθμούς στην κλίμακα ρίχτερ προκειμένου να υπάρχει ένα αξιόλογο δείγμα καταστροφών. Μετά την επεξεργασία των δεδομένων με χρήση pandas, η τελική μορφή του αρχείου περιέχει 413 γραμμές και 10 στήλες.



Εικόνα 11: USGS [11]

[11] https://earthquake.usgs.gov/

Πίνακας 9: USGS

Πηγή/Βάση δεδομένων	earthquake.usgs.gov (United States Geological Survey)
Περιγραφή	Το earthquake.usgs.gov είναι μία κρατική βάση η οποία περιέχει δεδομένα σεισμικών καταστροφών που συνέβησαν από το 1970 μέχρι σήμερα.
Είδη καταστροφών που περιλαμβάνονται	Περιλαμβάνει αποκλειστικά σεισμούς ανά την υφήλιο
Κατηγορίες των δεδομένων	Ημερομηνία έναρξης καταστροφής
	Γεωγραφικό μήκος και πλάτος που προσδιορίζουν τις συντεταγμένες του συμβάντος
	Τοποθεσία συμβάντος (Χώρα, Πόλη, Περιφέρεια).
	Τύπο καταστροφής
	Μοναδικό αναγνωριστικό
Λεπτομέρειες σχετικά με τα δεδομένα	Το αναγνωριστικό αντιστοιχεί σε μία συγκεκριμένη καταστροφή ως μοναδικό γνώρισμα.
	Δίνονται οι συντεταγμένες της τοποθεσίας, η χώρα και το αναγνωριστικό της, η περιοχή και η περιφέρεια αναζήτησης με ακρίβεια χιλιομέτρου.
	Οι πηγές που παρέχονται δίνουν παραπάνω αξιοπιστία στα δεδομένα.
	Υπάρχει μία περιεκτική και μία πιο αναλυτική περιγραφή της καταστροφής.
	Οι ανθρώπινες απώλειες και οι τραυματίες.
	Οι αλλαγές που επήλθαν στην μορφολογία του εδάφους (βύθιση, κλίση, μείωση μεγέθους κλπ.).
Σκοπός	Το earthquake.usgs.gov παρέχει αξιόπιστα δεδομένα καθώς παρέχονται μέσω του USGS, ενός

	αξιόπιστου κρατικού φορέα που διαθέτει τα δεδομένα του για αποφυγή μελλοντικών καταστροφών και ενημέρωση του κόσμου
Γλώσσα	Αγγλικά.
Δημόσια/Ιδιωτικά δεδομένα	Η διάθεση των δεδομένων της σελίδας είναι ελεύθερη προς όλους.
Type/format των δεδομένων	Αρχεία CSV, JSON.
Τρόποι εξαγωγής των δεδομένων	Είναι δυνατό για οποιονδήποτε δημόσιο χρήστη να κατεβάσει τα δεδομένα χωρίς να απαιτείται εγγραφή στον φορέα ,σε μορφή CSV, JSON και η βάση ανανεώνεται δυναμικά συνεχώς.

2.10 Καθορισμός κριτηρίων αξιολόγησης των πηγών/βάσεων δεδομένων

Σε αυτό το κεφάλαιο αναλύονται τα κριτήρια αξιολόγησης και επιλογής των συγκεκριμένων βάσεων δεδομένων. Στόχος είναι η βελτιστοποίηση της μεθόδου αναζήτησης των πηγών, χρησιμοποιώντας συγκεκριμένα κριτήρια αξιολόγησης τα οποία θα αναλυθούν παρακάτω.

2.11 Κριτήρια Αξιολόγησης

Πίνακας 10: Κριτήρια Αξιολόγησης

Τα κυριότερα κριτήρια είναι τα εξής:
1. Αξιολόγηση της εγκυρότητας των πηγών, με τη μέθοδο της διασταύρωσης των δεδομένων με άλλες αξιόπιστες πηγές και βάσεις.
2. Βασική προϋπόθεση είναι να υπάρχει επάρκεια δεδομένων που να καλύπτει όλο το φάσμα του αντικειμένου χωρίς να παραλείπονται καίριες πληροφορίες. Επίσης, να έχουν πρόσβαση όλοι οι χρήστες, χωρίς να απαιτείται συνδρομή, προκειμένου να είναι εύκολα προσβάσιμες οι πληροφορίες. Η ύπαρξη μεγάλου χρονικού εύρους διασφαλίζει την εφαρμογή συνθηκών σε ένα λογικό χρονικό διάστημα, προκειμένου τα συμπεράσματα που εξάγονται να είναι επαρκώς τεκμηριωμένα. Αναφέρεται το αντίκτυπο και οι συνέπειες της καταστροφής στους ανθρώπους, στα υλικά αγαθά και στη μορφολογία του εδάφους.
3. Η βάση να δίνει πληροφορίες για πολλές φυσικές καταστροφές και να μην περιορίζεται σε ένα συγκεκριμένο είδος.
4. Σημαντικό ρόλο στην επιλογή έχει το να μπορεί κανείς να χρησιμοποιήσει τις πληροφορίες που παρέχονται από τα δεδομένα, με στόχο να διευκολύνεται η πρόληψη αλλά και η πρόβλεψη μελλοντικών καταστροφών.

2.12 Επιλογή των πηγών/βάσεων προς αξιοποίηση

Πίνακας 11: Επιλογή Βάσεων

	Με βάση τα κριτήρια που ορίστηκαν στην προηγούμενη παράγραφο, καταλήγω στα εξής συμπεράσματα για τις δοθείσες πηγές:
Global Flood Database	Η βάση πληροί όλα τα προαναφερθέντα κριτήρια εκτός από το τρίτο. Αφορά μόνο πλημμύρες και ανεμοστρόβιλους. Άρα, αν και η δομή της βάσης είναι η επιθυμητή και τα δεδομένα που παρέχονται είναι επαρκή, το πλήθος των καταστροφών είναι πολύ περιορισμένο και δεν καλύπτει όλο το επιθυμητό φάσμα καταστροφών.
OurWorldInData	Η βάση OurWorldInData είναι η ιδανική, καθώς πληροί όλα τα κριτήρια σε αρκετά καλό βαθμό.

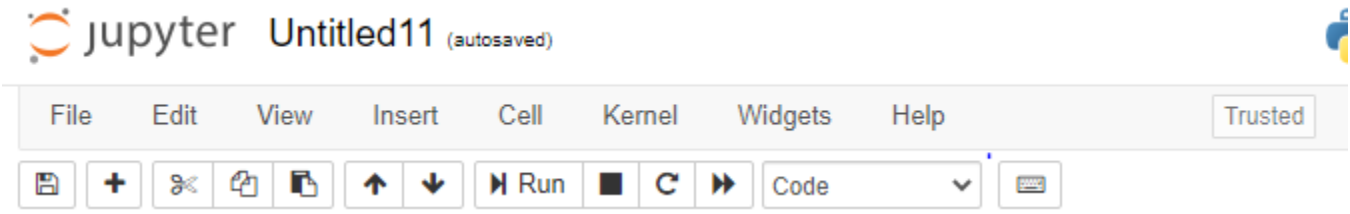
BDcatnat	Η τρίτη βάση έχει επίσης πληθώρα δεδομένων και είναι user friendly. Καλύπτει όλο το φάσμα των γνωστών καταστροφών με δυνατότητα εύκολου extract των δεδομένων. Το μόνο αρνητικό που έχει είναι το γεγονός ότι για να δει κάποιος ολόκληρη τη βάση απαιτείται δημιουργία premium χρήστη.
Copernicus	Στη συγκεκριμένη βάση, ενώ η δομή είναι η επιθυμητή αναφέρεται μόνο στις πυρκαγιές σε ευρωπαϊκές χώρες. Δεν παρέχει δεδομένα για τον υπόλοιπο κόσμο άρα δεν είναι η επιθυμητή.
FireserviceGR	Η βάση FireserviceGR είναι αρκετά ελλιπής καθώς τα δεδομένα της περιορίζονται σε πυρκαγιές και περιορίζονται στον Ελλαδικό χώρο. Παρέχει όμως λεπτομερή στοιχεία για την κινητοποίηση των πυροσβεστικών δυνάμεων καθώς και τα θύματα και τους τραυματίες της πυρκαγιάς. Άρα πληροί, εν μέρει, τα κριτήρια που τέθηκαν.
Global Internal displacement Database	Η έκτη βάση είναι αρκετά πλήρης καθώς πέραν του γεγονότος ότι καλύπτει τις περισσότερες καταγεγραμμένες καταστροφές, έχει πρόσθετες πληροφορίες για τις μετατοπίσεις του πληθυσμού που αυτές δημιούργησαν. Επίσης έχει τεράστιο πλήθος καταχωρήσεων άρα και μεγαλύτερη αξιοπιστία.
EprecSilence	Η έβδομη βάση είναι η ιδανική καθώς προσφέρει εύκολη πρόσβαση στα δεδομένα για όλους τους χρήστες.
Global Landslide Catalog	Η συγκεκριμένη βάση πέρα από την αξιοπιστία που προσφέρει ο πάροχος των δεδομένων (NASA), παρέχει και μεγάλη πληθώρα αυτών (πάνω από 11.000 καταχωρήσεις). Επίσης περιέχει πολλά είδη καταστροφών και τα δεδομένα είναι ελεύθερα προς διάθεση για όλους.
United States Geological Survey	Η βάση 9 αν και περιέχει αποκλειστικά και μόνο σεισμούς και όχι άλλες φυσικές καταστροφές περιέχει πληροφορίες για την ένταση της καταστροφής (κλίμακα ρίχτερ) και τις επιπτώσεις που αυτή είχε. Περιέχει επίσης πληροφορίες για την τοποθεσία και την ημερομηνία που συνέβη. Πληροί, λοιπόν, τα κριτήρια που αναφέρθηκαν στην προηγούμενη παράγραφο.

Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

3 Ανάλυση τεχνολογιών που χρησιμοποιήθηκαν

3.1 Jupyter Notebook

Το Jupyter Notebook[10] αποτελεί μια διαδικτυακή διαδραστική πλατφόρμα υπολογισμών για κώδικα και δεδομένα. Το όνομα Jupyter προκύπτει από την ένωση των σημαντικότερων γλωσσών προγραμματισμού που υποστηρίζει (Julia, Python και R.Jupyter). Συνολικά υποστηρίζει περισσότερες από 40 γλώσσες προγραμματισμού, στη συγκεκριμένη εργασία έγινε χρήση της python. Συγκεντρώνει όλα τα δεδομένα ενός project, για να γίνεται πιο εύκολη η επεξεργασία, ο συντονισμός και η προβολή τους. Ανοίγοντας το αρχείο με χρήση του file-open ξεκινάνε οι εντολές από το πρώτο κελί (In[1]) και εμφανίζεται το output στο αντίστοιχο κελί (Out[1]). Χρησιμοποιήθηκαν αρχεία excel και csv στα οποία έγιναν τροποποιήσεις με σκοπό να γίνουν extract τα δεδομένα μέσω αρχείου JSON.



Εικόνα 12: JUPYTER [12]

3.2 NumPy(Numerical Python)

Το NumPy [11](Numerical python) αποτελεί ένα πακέτο επέκτασης της Python για επιστημονικούς υπολογισμούς. Επίσης, παρέχει προχωρημένες δυνατότητες χειρισμού πινάκων N-διαστάσεων. Η βασική διαφορά της με τις λίστες της Python είναι το γεγονός ότι, σε αντίθεση με αυτήν, έχει συγκεκριμένο μέγεθος που ορίζεται κατά την δημιουργία της. Αυτοί οι ορισμένοι πίνακες αποτελούν πολυδιάστατες συλλογές ομοιόμορφων στοιχείων και χρησιμοποιούνται πολύ συχνά στην επιστήμη δεδομένων, όπου η ταχύτητα και οι πόροι έχουν μεγάλη σημασία. Παρέχει έναν πίνακα έως και 50 φορές ταχύτερο από τις παραδοσιακές λίστες python. Είναι open source και χρησιμοποιείται δωρεάν.

```
import pandas as pd
import numpy as np
```

Εικόνα 13: NUMPY

[12] https://jupyter.org/

3.3 Pandas

Η Pandas[12] είναι βιβλιοθήκη ανάλυσης δεδομένων της Python. Χρησιμοποιείται κυρίως για διαλογή και εκκαθάριση δεδομένων κυρίως csv και excel, προκειμένου να γίνει αξιοποίησή τους και εξαγωγή χρήσιμων πληροφοριών. Αποτελεί επέκταση της βιβλιοθήκης NumPy της οποίας οι λειτουργίες επεκτείνονται για περαιτέρω ανάλυση. Χαρακτηριστικό της είναι η ευκολία και η ταχύτητα στη διαχείριση των εισαχθέντων δεδομένων.
Για την εκπόνηση της διπλωματικής άσκησης έγινε χρήση pandas για την δημιουργία και την τροποποίηση των Disaster Databases και των Dataframes. Η διαδικασία που ακολουθήθηκε στα πεδία αναλύονται παρακάτω:

1) Εισαγωγή pandas και NumPy, προκειμένου να γίνει εφικτή η χρησιμοποίησή τους στα επόμενα βήματα.
2) Τα αρχεία που χρησιμοποιήθηκαν ήταν της μορφής excel ή csv.
3) Με την χρήση του df ανοίγει το αρχείο που έχει επιλεγεί και το βλέπουμε σε μορφή dataframework.

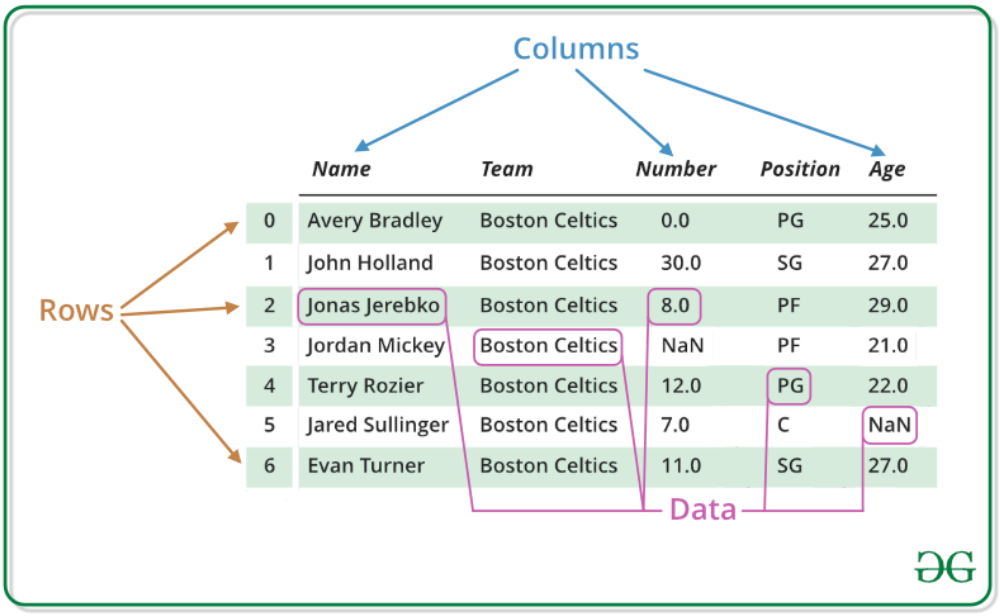
Κάποιες βάσεις που περιείχαν δεδομένα και columns, έπρεπε να εξαλειφτούν ή να γίνει αλλαγή της επικεφαλίδας τους. Η βιβλιοθήκη παρέχει πληθώρα εντολών για το σκοπό αυτό.

```
In [2]: import pandas as pd
```

Εικόνα 14: Κλήση PANDAS, alias as PD

```
In [5]: df = pd.read_csv(r'C:\Users\vigsa\Downloads\query (1).csv')
```

Εικόνα 15: Κλήση συγκεκριμένου αρχείου csv μέσω μονοπατιού και ανάθεσή του στο Dataframe



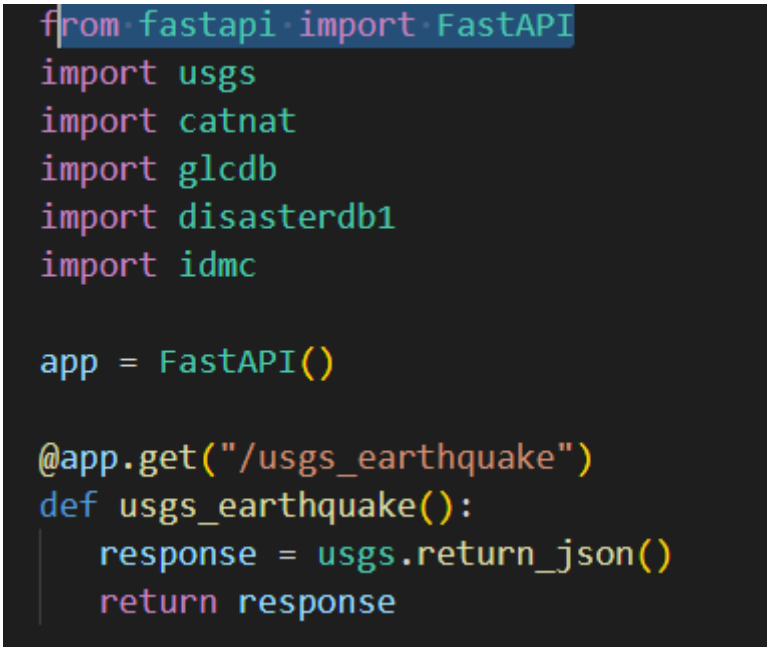
Εικόνα 16: PANDAS [13]

Παράδειγμα αρχείου που επεξεργάζονται τα Pandas

3.4 FastAPI

Το FastAPI[13] είναι ένα framework της python και μία ομάδα εργαλείων που δίνει τη δυνατότητα να καλούνται συναρτήσεις μέσω μίας διασύνδεσης REST για την υλοποίηση ποικίλλων εφαρμογών. Για να αποκτηθεί πρόσβαση χρησιμοποιείται ένα REST API με την οποία καλούνται δομικά στοιχεία μίας εφαρμογής. Πρέπει να γίνει import χρησιμοποιώντας το from fastapi import FastAPI στην main. Με την @app.get δέχεται η main.py τα δεδομένα από την usgs και επιστρέφει το JSON αρχείο κάνοντας χρήση του endpoint.

[13] https://www.geeksforgeeks.org/python-pandas-dataframe/



```python
from fastapi import FastAPI
import usgs
import catnat
import glcdb
import disasterdb1
import idmc

app = FastAPI()

@app.get("/usgs_earthquake")
def usgs_earthquake():
    response = usgs.return_json()
    return response
```

Εικόνα 17: FASTAPI

Ένα σημείο του API ονομάζεται endpoint, στο οποίο το FastAPI συνδέεται με το πρόγραμμα λογισμικού. Το FastAPI λειτουργεί στέλνοντας αιτήματα για πληροφορίες από έναν web server και λαμβάνει μία απάντηση.

4 Τύποι δεδομένων που αναλύονται στις βάσεις

Οι βάσεις που μελετήθηκαν περιείχαν ποικίλα δεδομένα που αφορούν φυσικές καταστροφές. Προκειμένου να ακολουθήσουν ένα επιθυμητό μοτίβο έγιναν διάφορες τροποποιήσεις και απαλοιφές στηλών με την χρήση python pandas. Τα χαρακτηριστικά που θα μελετηθούν φαίνονται παρακάτω με τη μορφή στηλών και καταχωρήσεων σε αυτές σε μορφή JSON. Έπρεπε οι 6 βάσεις να έχουν κοινά χαρακτηριστικά προς μελέτη εφ’ όσον αυτά υπήρχαν ως καταχωρήσεις χωρίς να προστεθούν τεχνητά. Αυτά αναλύονται εκτενέστερα στη συνέχεια σε μορφή πίνακα με πεδία:
Requested Field (όνομα της στήλης),
Field type (τύπος δεδομένων που περιέχει)
Field Description (περιγραφή πεδίου)

4.1 Εμφάνιση δεδομένων που προκύπτουν από την χρήση JSON


```
[
  {
    "startyear": 1864,
    "location": "The Great Sheffield Floods",
    "country": "UK",
    "disaster_type": "Floods",
    "casualties": 240,
    "injuredPeopleNumber": null,
    "database": "Disaster-database-UP-TO-DATE_1"
  },
  {
    "startyear": 1902,
    "location": "Ibrox Stadium, Glasgow, Scotland",
    "country": "UK",
    "disaster_type": "Crowd",
    "casualties": 25,
    "injuredPeopleNumber": 537,
    "database": "Disaster-database-UP-TO-DATE_1"
  },
  {
    "startyear": 1903,
    "location": "Iroquois Theatre Fire",
    "country": "USA",
    "disaster_type": "Crowd / Fire",
    "casualties": 602,
    "injuredPeopleNumber": "250+",
    "database": "Disaster-database-UP-TO-DATE_1"
  },
  {
    "startyear": 1905,
    "location": "Edinburgh Empire Palace Theatre Fire",
    "country": "UK",
    "disaster_type": "Fire",
    "casualties": 11,
    "injuredPeopleNumber": 0,
    "database": "Disaster-database-UP-TO-DATE_1"
  },
```


Εικόνα 18: DISASTER DATABASE UP TO DATE

```
[
  {
    "country": "Abyei Area",
    "StartDate": "2018-07-01",
    "event_description": "Abyie: Flood - 01/07/2018",
    "disaster_category": "Weather related",
    "disaster_Type": "Flood",
    "Internal Displacements": 2,
    "database": "IDMC_Internal_Displacement",
    "startYear": "2018",
    "startMonth": "07",
    "startDay": "01"
  },
  {
    "country": "Abyei Area",
    "StartDate": "2019-06-01",
    "event_description": "Abyei: Flood - southern parts - 01/06/2019",
    "disaster_category": "Weather related",
    "disaster_Type": "Flood",
    "Internal Displacements": 40000,
    "database": "IDMC_Internal_Displacement",
    "startYear": "2019",
    "startMonth": "06",
    "startDay": "01"
  },
  {
    "country": "Afghanistan",
    "StartDate": "2008-01-01",
    "event_description": null,
    "disaster_category": "Weather related",
    "disaster_Type": "Extreme temperature",
    "Internal Displacements": null,
    "database": "IDMC_Internal_Displacement",
    "startYear": "2008",
    "startMonth": "01",
    "startDay": "01"
  },
```


Εικόνα 19: IDMC

```
[
  {
    "country": " Australia",
    " latitude": -37.9681,
    " longitude": 145.7093,
    "disaster_Type": " 06 Forest fires",
    "startyear": 2019,
    "startmonth": " March",
    " Start date": " 2019-03-01",
    " end date": " 2019-03-08",
    "injuredPeopleNumber": 0,
    "casualties": 0,
    "database": "BD Catnat_exemple",
    "Endyear": " 201",
    "EndDay": "-08",
    "EndMonth": "-0",
    "StartYear": " 201",
    "StartDay": "-01",
    "StartMonth": "-0"
  },
  {
    "country": "Afghanistan",
    " latitude": 31.6288,
    " longitude": 65.7371,
    "disaster_Type": " 01 Floods and mudslides",
    "startyear": 2019,
    "startmonth": " March",
    " Start date": " 2019-03-01",
    " end date": " 2019-03-02",
    "injuredPeopleNumber": 20,
    "casualties": 83,
    "database": "BD Catnat_exemple",
    "Endyear": " 201",
    "EndDay": "-02",
    "EndMonth": "-0",
    "StartYear": " 201",
    "StartDay": "-01",
    "StartMonth": "-0"
```


Εικόνα 20: BDCATNAT

```
[
  {
    "event_Description": "occurred early in morning, 11 villagers buried in 7 houses",
    "location": "Sigou Village, Loufan County, Shanxi Province",
    "disaster_Size": "large",
    "casualties": "11",
    "injuredPeopleNumber": null,
    "country": "China",
    "longitude": 107.45,
    "latitude": 32.5625,
    "database": "Global Landslide Catalog",
    "startYear": "2008",
    "startMonth": "01",
    "startDay": "08",
    "startTime": "12:00:00 AM",
    "startDate": "08/01/2008"
  },
  {
    "event_Description": "Hours of heavy rain are to blame for an overnight mudslide in Lake Oswego. ",
    "location": "Lake Oswego, Oregon",
    "disaster_Size": "small",
    "casualties": "0",
    "injuredPeopleNumber": null,
    "country": "United States",
    "longitude": -122.663,
    "latitude": 45.42,
    "database": "Global Landslide Catalog",
    "startYear": "2009",
    "startMonth": "02",
    "startDay": "01",
    "startTime": "02:00:00 AM",
    "startDate": "01/02/2009"
  },
  {
    "event_Description": "(CBS/AP) At least 10 people died and as many as 80 were still missing Wednesday in
central Peru after torrential rains swelled three rivers, forcing them over their banks and causing devastating
landslides earlier in the week. ",
    "location": "San Ramon district, 195 miles northeast of the capital, Lima, ",
    "disaster_Size": "large",
    "casualties": "10",
    "injuredPeopleNumber": null,
```


Εικόνα 21: GLC

```
[
  {
    "time": "2004-12-29T05:56:47.540Z",
    "latitude": 8.791,
    "longitude": 93.198,
    "mag": 6.2,
    "location": "Nicobar Islands, India region",
    "disaster_type": "earthquake",
    "database": "earthquake.usgs.gov",
    "startYear": "2004",
    "startMonth": "12"
  },
  {
    "time": "2004-12-29T01:50:52.570Z",
    "latitude": 9.109,
    "longitude": 93.756,
    "mag": 6.1,
    "location": "Nicobar Islands, India region",
    "disaster_type": "earthquake",
    "database": "earthquake.usgs.gov",
    "startYear": "2004",
    "startMonth": "12"
  },
  {
    "time": "2004-12-27T09:39:06.800Z",
    "latitude": 5.348,
    "longitude": 94.65,
    "mag": 6.1,
    "location": "78 km WSW of Banda Aceh, Indonesia",
    "disaster_type": "earthquake",
    "database": "earthquake.usgs.gov",
    "startYear": "2004",
    "startMonth": "12"
  },
  {
    "time": "2004-12-26T19:19:55.570Z",
    "latitude": 2.794,
    "longitude": 94.162,
```


Εικόνα 22: USGS

```
[
  {
    "startDate": 1483920000000,
    "disasterSize": "ΜΙΚΡΗ",
    "carForce": 1.0,
    "manpowerForce": 3.0,
    "vehiclesForce": 0.0,
    "injuredpeoplenumber": 0.0,
    "casualties": 0.0,
    "Καταστροφές": 0.0,
    "database": "FireServiceGR"
  },
  {
    "startDate": 1485561600000,
    "disasterSize": "ΜΕΣΑΙΑ",
    "carForce": 2.0,
    "manpowerForce": 5.0,
    "vehiclesForce": 0.0,
    "injuredpeoplenumber": 0.0,
    "casualties": 0.0,
    "Καταστροφές": 0.0,
    "database": "FireServiceGR"
  },
  {
    "startDate": 1485820800000,
    "disasterSize": "ΜΕΣΑΙΑ",
    "carForce": 1.0,
    "manpowerForce": 2.0,
    "vehiclesForce": 0.0,
    "injuredpeoplenumber": 0.0,
    "casualties": 0.0,
    "Καταστροφές": 0.0,
    "database": "FireServiceGR"
  },
  {
    "startDate": 1486080000000,
    "disasterSize": "ΜΙΚΡΗ",
    "carForce": 2.0,
    "manpowerForce": 6.0,
    "vehiclesForce": 0.0,
    "injuredpeoplenumber": 0.0,
    "casualties": 0.0,
    "Καταστροφές": 0.0,
    "database": "FireServiceGR"
  },
  {
    "startDate": 1486425600000,
```


Εικόνα 23: FireserviceGR

4.2 Αναλυτική Περιγραφή και τύπος δεδομένων:

Πίνακας 12: Περιγραφή Δεδομένων Βάσεων

Requested Field	Field_type	Field_Description
Country	`object`	Χώρα στην οποία συνέβη η καταστροφή
location	`object`	Τοποθεσία
Event_description	`object`	Αναλυτική περιγραφή γεγονότος η οποία μπορεί να περιλαμβάνει ημερομηνίες και πληγέντες
disaster_Size		Δείχνει Μέγεθος καταστροφής σε small, medium, large size
Disaster_category	`object`	Τύπος καταστροφής που βασίζεται στην προέλευσή της (σχετιζόμενη με τον καιρό ή γεωφυσική)
Disaster_type	`object`	Ανάλογα με τον τύπο της καταστροφής δημιουργούνται υποκατηγορίες για γεωφυσικές και σχετιζόμενες με τον καιρό αντίστοιχα
latitude	`float64`	Το γεωγραφικό πλάτος που ορίζει μαζί με το γεωγραφικό μήκος τις συντεταγμένες που έλαβε χώρα η καταστροφή
Longitude	`float64`	Το γεωγραφικό της μήκος
mag	`float64`	Δείχνει το μέγεθος ενός σεισμού στη μονάδα μέτρησης ρίχτερ
casualties	`int64(BD)`	Οι ανθρώπινες απώλειες που καταμετρήθηκαν κατά τη διάρκεια της καταστροφής

injuredPeopleNumber	`float64, int64(BD)`	Ο αριθμός των τραυματιών που καταμετρήθηκε
Internal_Displacements	`Object, int64(BD)`	Οι άνθρωποι που αναγκάστηκαν να μετοικήσουν (μόνιμα ή προσωρινά) λόγω των καταστροφών που επήλθαν στα σπίτια τους
startDate	`object`	Η ημερολογιακή ημέρα έναρξής της (1-30)
startTime	`object`	Ο χρόνος που ξεκίνησε
startYear	`object`	Το ημερολογιακό έτος έναρξης (μορφής π.χ. 2022)
startMonth	`object`	Ο μήνας που άρχισε (1-12)
sstartDay	`object`	Η ημέρα έναρξής της (1-30)
endYear	`object`	Το ημερολογιακό έτος λήξης της καταστροφής (μορφής π.χ. 2022)
endMonth	`object`	Ο μήνας που τελείωσε(1-12)
endDay	`object`	Η ημέρα λήξης της (1-30)
database	`object`	Το όνομα της βάσης δεδομένων από την οποία προήλθαν τα δεδομένα καταστροφής

5 Ανάλυση Αλγορίθμων Μηχανικής Μάθησης

5.1 Αλγόριθμοι μηχανικής Μάθησης(ML Algorithms)

Οι αλγόριθμοι μηχανικής μάθησης [14](ML algorithms) είναι η μέθοδος που χρησιμοποιεί ένα AI σύστημα για να προβλέπει τα δεδομένα που εξάγονται ενώ δέχεται συγκεκριμένο input. Χρησιμοποιούνται κυρίως για ταξινόμηση και παλινδρόμηση δεδομένων. Χωρίζονται σε δύο υποκατηγορίες supervised και unsupervised. Οι supervised αλγόριθμοι δέχονται και εξάγουν δεδομένα τα οποία τους παρέχονται μέσω τιτλοφόρησης (labeling). Οι unsupervised αλγόριθμοι έχουν μεγαλύτερη ελευθερία στα δεδομένα που επεξεργάζονται, καθώς δεν χρειάζεται να είναι τιτλοφορημένα ή ταξινομημένα. Ως ταξινόμηση ορίζεται η διαδικασία κατά την οποία αναγνωρίζονται και ομαδοποιούνται αντικείμενα και δεδομένα σε κατηγορίες με βάση κάποιο κοινό χαρακτηριστικό. Έμφαση θα δοθεί στους αλγόριθμους ταξινόμησης (classification algorithms) οι οποίοι χρησιμοποιούνται για να προβλέψουν την έκβαση διαδικασιών με βάση πιθανολογικά στοιχεία. Δέχονται ως όρισμα κάποια δεδομένα που οδήγησαν σε συγκεκριμένη έκβαση ενός προβλήματος και με βάση αυτά προβλέπουν την έκβαση ενός άλλου προβλήματος με παρεμφερή χαρακτηριστικά. Επίσης δημιουργούν κάποιο μοτίβο αναγνώρισης δεδομένων για να χρησιμοποιηθεί σε μελλοντικές καταστάσεις.

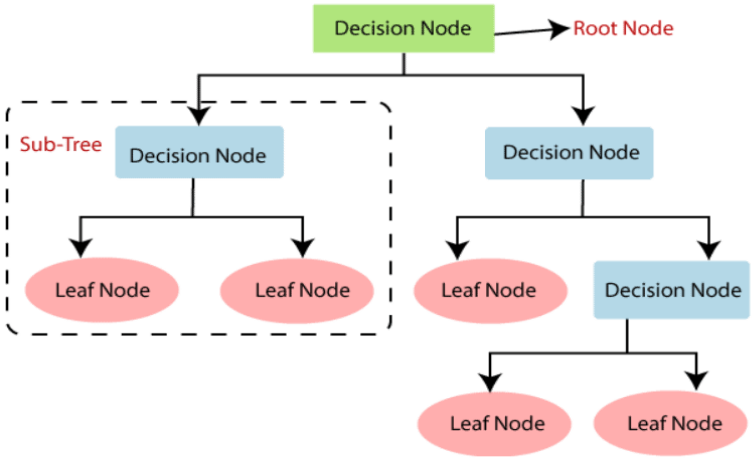
Κυριότερα είδη αλγορίθμων ταξινόμησης:

1)Logistic regression
2)Decision tree
3)K-Nearest Neighbours
4)SVM
5)Naïve Bayes Classifier

5.1.1 Decision Tree

Το δέντρο απόφασης [15] είναι ένας πολύ δημοφιλής αλγόριθμος μηχανικής μάθησης. Λειτουργεί τόσο για γραμμικά όσο και για μη γραμμικά δεδομένα. Μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση. Με τις μεγάλες βιβλιοθήκες και τα πακέτα που είναι διαθέσιμα στην Python και την R, είναι εύκολα προσβάσιμα από όλους. Λειτουργεί με βάση την απόφαση σχετικά με τις συνθήκες των χαρακτηριστικών. Οι κόμβοι είναι οι δοκιμές ενός

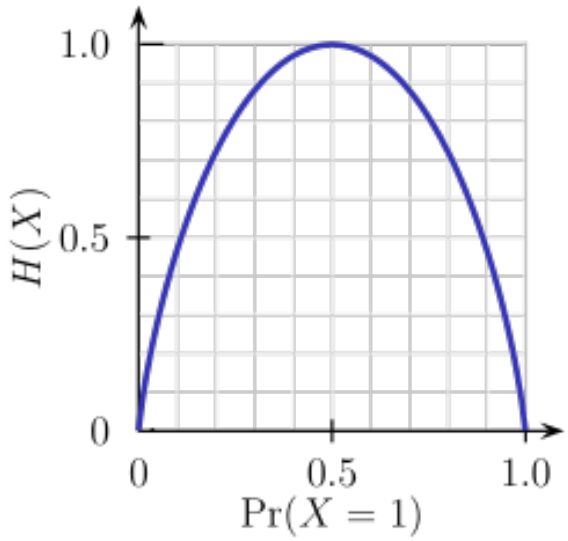
χαρακτηριστικού, ο κλάδος αντιπροσωπεύει το αποτέλεσμα των δοκιμών και οι κόμβοι φύλλων είναι οι αποφάσεις με βάση τις συνθήκες.



Εικόνα 24: Decision Tree [14]

Ο ριζικός κόμβος ή το πρώτο χαρακτηριστικό δοκιμής επιλέγεται με βάση ένα στατιστικό μέτρο που ονομάζεται information gain.
Συνολικά, η επιλογή των χαρακτηριστικών δοκιμής εξαρτάται από το "μέτρο καθαρότητας". Αυτά τα μέτρα καθαρότητας είναι: Information gain, Gain Ratio, Gini index.
Το κέρδος πληροφορίας βοηθά στη μέτρηση της μείωσης της αβεβαιότητας ενός συγκεκριμένου χαρακτηριστικού. Βοηθά επίσης να αποφασιστεί ποιο χαρακτηριστικό είναι καλό ως ριζικός κόμβος.
Η εντροπία είναι ένα μέτρο της τυχαιότητας των πληροφοριών που υποβάλλονται σε επεξεργασία. Όσο υψηλότερη η εντροπία, τόσο πιο δύσκολο είναι να εξαχθούν συμπεράσματα από τις πληροφορίες αυτές (τύπος εντροπίας). Από το παρακάτω διάγραμμα είναι προφανές πως όταν η εντροπία H(X) είναι όταν η πιθανότητα είναι από το 0 ως το 1.

[14] https://k21academy.com/datascience/decision-tree-algorithm/



Εικόνα 25: Εντροπία [15]

Πίνακας 13: Decision Tree Pros and Cons

Πλεονεκτήματα	**Μειονεκτήματα**
Χρησιμοποιείται για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης.	Όσο αυξάνονται οι εγγραφές τόσο αυξάνεται ο χρόνος για την εκπαίδευση του δέντρου απόφασης σε αριθμητικές μεταβλητές.
Ταξινόμηση μη γραφικών δεδομένων.	Όταν έχουμε δέντρο απόφασης με πολλά χαρακτηριστικά είναι πολύ πιο χρονοβόρο.
Πολύ γρήγορος και αποτελεσματικός σε σύγκριση με τον KNN και άλλους αλγορίθμους ταξινόμησης.	Το πρόβλημα της υπερπροσαρμογής μπορεί να επιλυθεί θέτοντας περιορισμούς στις παραμέτρους του μοντέλου και της μεθόδου κλαδέματος.
Εύκολη ερμήνευση αποτελεσμάτων.	Εάν το μέγεθος των δεδομένων είναι πολύ μεγάλο, τότε ένα μόνο δέντρο μπορεί να αναπτύξει πολλούς κόμβους, γεγονός που μπορεί να οδηγήσει σε πολυπλοκότητα και να οδηγήσει σε υπερπροσαρμογή.
Χρήση οποιουδήποτε τύπου δεδομένων.	Στο πρόβλημα της υπερπροσαρμογής,

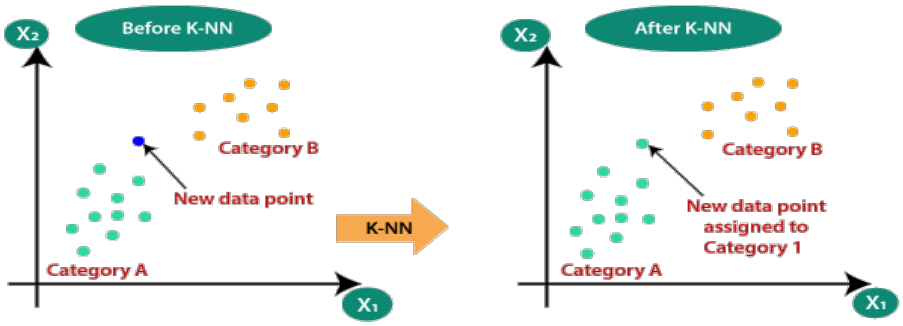
[15] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

	υπάρχει πολύ μεγάλη διακύμανση στην έξοδο, η οποία οδηγεί σε πολλά σφάλματα στην τελική εκτίμηση και μπορεί να εμφανίσει μεγάλη ανακρίβεια
Χρήσιμος στην εξερεύνηση δεδομένων.	
Τα δέντρα αποφάσεων έχουν καλύτερη ισχύ με την οποία μπορούμε να δημιουργήσουμε νέες μεταβλητές/ χαρακτηριστικά για τη μεταβλητή του αποτελέσματος.	
Σύντομη προετοιμασία δεδομένων.	
Το δέντρο αποφάσεων είναι μη παραμετρικό*	

***Ως μη παραμετρική μέθοδος ορίζεται η μέθοδος στην οποία δεν υπάρχουν υποθέσεις σχετικά με τη χωρική κατανομή και τη δομή του ταξινομητή.**

5.1.2 K- Nearest Neighbour

Ο αλγόριθμος KNN [16] χρησιμοποιείται κυρίως ως ταξινομητής, είναι εύκολα κατανοητός και απλός στην εφαρμογή. Είναι ιδανικός σε περιπτώσεις όπου τα σημεία δεδομένων είναι καλά καθορισμένα ή μη γραμμικά.



Εικόνα 26: K-Nearest Neighbour [16]

Εάν η τιμή του K=1, τότε θα χρησιμοποιήσουμε μόνο τον πλησιέστερο γείτονα για να καθορίσουμε την κλάση ενός σημείου δεδομένων. Σε αυτή την περίπτωση, το σημείο δεδομένων X ανήκει στην ομάδα A, καθώς ο πλησιέστερος γείτονάς του βρίσκεται στην ίδια ομάδα.

16 https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Εάν η τιμή του K=10, τότε θα χρησιμοποιήσουμε τους δέκα πλησιέστερους γείτονες. Θεωρούμε ένα μη ταξινομημένο σημείο δεδομένων X. Εάν η ομάδα A έχει περισσότερα από δέκα σημεία δεδομένων και η τιμή του K είναι ίση με 10, τότε το σημείο δεδομένων X εξακολουθεί να ανήκει στην ομάδα A, καθώς όλοι οι πλησιέστεροι γείτονές του βρίσκονται στην ίδια ομάδα.
Το γεγονός ότι ο ταξινομητής αναθέτει την κατηγορία με τον μεγαλύτερο αριθμό ψήφων ισχύει ανεξάρτητα από τον αριθμό των κατηγοριών που υπάρχουν.
Υπάρχουν τέσσερις τρόποι υπολογισμού του μέτρου απόστασης μεταξύ του σημείου δεδομένων και του πλησιέστερου γείτονά του: Ευκλείδεια απόσταση, απόσταση Μανχάταν, απόσταση Χάμινγκ και απόσταση Μινκόφσκι.
Από τις τρεις, η ευκλείδεια απόσταση είναι η πιο συχνά χρησιμοποιούμενη συνάρτηση ή μετρική απόσταση. Για την υλοποίηση του αλγορίθμου KNN χρησιμοποιούνται γλώσσες προγραμματισμού όπως η Python και η R.
Για να επικυρωθεί η ακρίβεια της ταξινόμησης KNN, χρησιμοποιείται ένας πίνακας σύγχυσης. Για την επικύρωση χρησιμοποιούνται επίσης άλλες στατιστικές μέθοδοι, όπως ο έλεγχος του λόγου πιθανοφάνειας (likelihood-ratio test).

Πίνακας 14: K-Nearest Neighbour

Για κάθε σημείο δεδομένων στα δεδομένα:
1. Βρίσκουμε την ευκλείδεια απόσταση από όλα τα δείγματα δεδομένων εκπαίδευσης.
2. Αποθήκευση των αποστάσεων σε διατεταγμένη λίστα και ταξινόμησή της.
3. Επιλογή των κορυφαίων K καταχωρίσεων από τον ταξινομημένο κατάλογο.
4. Ετικέτα του σημείου δοκιμής με βάση την πλειοψηφία των κλάσεων που υπάρχουν στα επιλεγμένα σημεία.

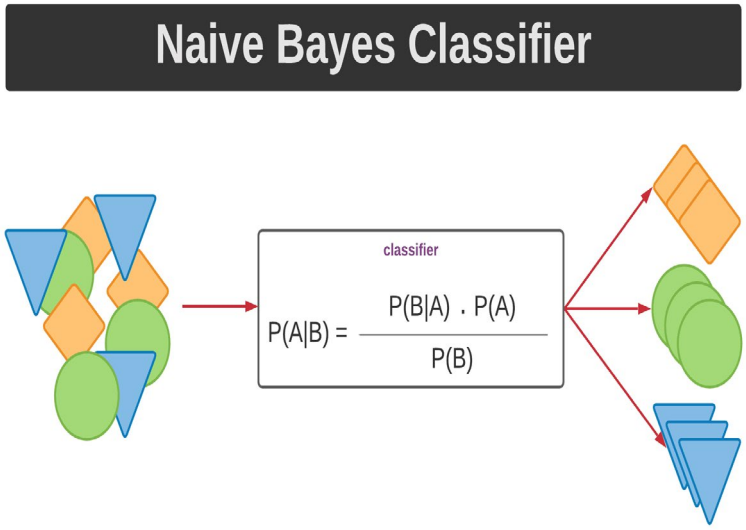
Γιατί να χρησιμοποιούμε τον αλγόριθμο KNN;
Ο KNN είναι ένας από τους παλαιότερους αλλά ακριβείς αλγορίθμους που χρησιμοποιούνται για μοντέλα ταξινόμησης προτύπων και παλινδρόμησης. Ένα από τα σημαντικότερα πλεονεκτήματα της χρήσης του αλγορίθμου KNN είναι ότι δεν χρειάζεται να δημιουργηθεί κάποιο μοντέλο ή να ρυθμιστούν διάφορες παράμετροι. Δεδομένου ότι ο προγνωστικός αλγόριθμος KNN υπολογίζει τα πάντα από την αρχή, ενδέχεται να μην είναι ιδανικός για μεγάλα σύνολα δεδομένων.

Πίνακας 15: K-Near Algorithm Pros and Cons

Πλεονεκτήματα	**Μειονεκτήματα**
Είναι εύκολος στην κατανόηση και απλός στην εφαρμογή.	Το σχετικό υπολογιστικό κόστος είναι υψηλό, καθώς αποθηκεύει όλα τα δεδομένα εκπαίδευσης.

Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης	Απαιτεί υψηλή αποθήκευση στη μνήμη.
Είναι ιδανικός για μη γραμμικά δεδομένα, καθώς δεν υπάρχει καμία υπόθεση σχετικά με τα υποκείμενα δεδομένα.	Ανάγκη προσδιορισμού της τιμής του K.
Μπορεί φυσικά να χειριστεί περιπτώσεις πολλαπλών κλάσεων.	Η πρόβλεψη είναι αργή εάν η τιμή του N είναι υψηλή.
Μπορεί να αποδώσει καλά με αρκετά αντιπροσωπευτικά δεδομένα.	Ευαισθησία σε άσχετα χαρακτηριστικά.

5.1.3 Naïve Bayes Classifier Algorithm



Εικόνα 27: Naïve Bayes Classifier Algorithm [17]

[17] https://jagan-singhh.medium.com/naive-bayes-classifier-99e3e618f8db

Θεώρημα του Bayes:
Ο αλγόριθμος του Bayes [17] βασίζεται στο γνωστό θεώρημά του που δημοσιεύτηκε το 1763 και έδινε τον τύπο υπολογισμού της δεσμευμένης πιθανότητας, δηλαδή την πιθανότητα ενός ενδεχομένου με δεδομένο το ότι ένα άλλο ενδεχόμενο έχει συμβεί.
Ο τύπος για το θεώρημα του Bayes δίνεται ως εξής:
Αλγόριθμος ταξινομητή Naïve Bayes:
Οπου:

P(A|B) is Posterior probability: Πιθανότητα της υπόθεσης A για το παρατηρούμενο γεγονός B.

P(B|A) is Likelihood probability: : Πιθανότητα της απόδειξης δεδομένου ότι η πιθανότητα μιας υπόθεσης είναι αληθής.

P(A) is Prior Probability: Πιθανότητα της υπόθεσης πριν από την παρατήρηση των αποδεικτικών στοιχείων.

P(B) is Marginal Probability: Πιθανότητα της απόδειξης.

Για τη δημιουργία ενός μοντέλου Naive Bayes σε Python και R, χρησιμοποιειται το scikit learn (βιβλιοθήκη python).

Υπάρχουν τρεις βασικοί τύποι μοντέλων Naive Bayes:

1. Gaussian: Χρησιμοποιείται στην ταξινόμηση και υποθέτει ότι τα χαρακτηριστικά ακολουθούν κανονική κατανομή.

2. Πολυωνυμική: Χρησιμοποιείται για διακριτές μετρήσεις. Για παράδειγμα, ας πούμε ότι έχουμε ένα πρόβλημα ταξινόμησης κειμένου. Εδώ μπορούμε να θεωρήσουμε δοκιμές Bernoulli που είναι ένα βήμα παραπέρα και αντί για "λέξη που εμφανίζεται στο έγγραφο", έχουμε "μετρήστε πόσο συχνά εμφανίζεται η λέξη στο έγγραφο", μπορείτε να το σκεφτείτε ως "αριθμός των φορών που παρατηρείται ο αριθμός αποτελέσματος x_i κατά τη διάρκεια των n δοκιμών".

3. Bernoulli: Το διωνυμικό μοντέλο είναι χρήσιμο εάν τα διανύσματα των χαρακτηριστικών σας είναι δυαδικά (δηλαδή μηδενικά και μονάδες). Μια εφαρμογή θα ήταν η ταξινόμηση κειμένου με το μοντέλο "σακούλα λέξεων" όπου τα 1 & 0 είναι "η λέξη εμφανίζεται στο έγγραφο" και "η λέξη δεν εμφανίζεται στο έγγραφο" αντίστοιχα.

Οι εφαρμογές του:
Ο αλγόριθμος είναι γρήγορος, απλός και χρησιμοποιείται για εργασίες ταξινόμησης κειμένου. Δεδομένου ότι μπορεί να χρησιμοποιηθεί και για ταξινόμηση πολλαπλών κατηγοριών, θεωρείται ένας πολύ ευέλικτος και

ευέλικτος ταξινομητής. Η πλειονότητα των ερευνητικών εργασιών σχετικά με την ταξινόμηση κειμένου ξεκινά με τη χρήση του ταξινομητή Naive Bayes ως βασικού μοντέλου.

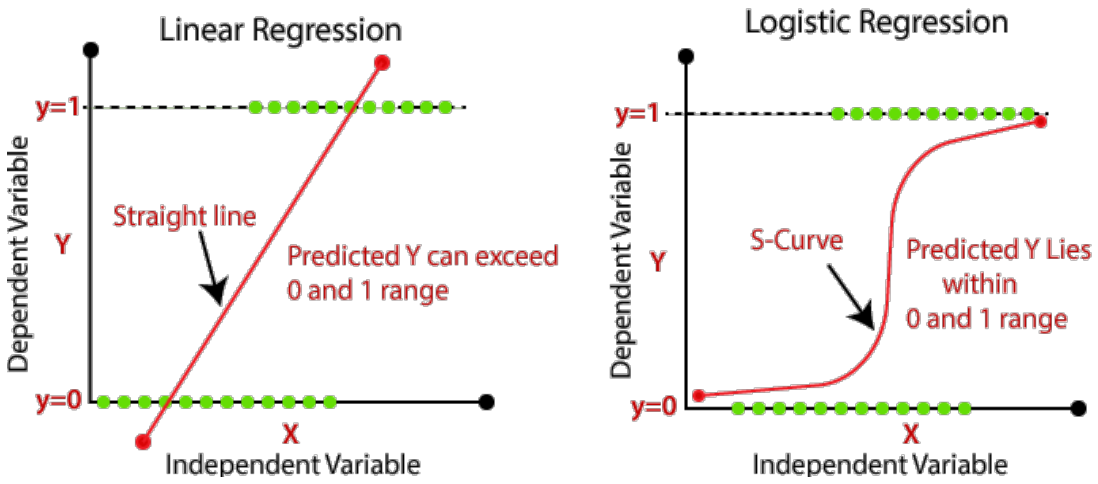
Πίνακας 16: Naïve Bayes Classifier Algorithm Pros and Cons

Πλεονεκτήματα	Μειονεκτήματα
Είναι εύκολο και γρήγορο να προβλέψει την κλάση του συνόλου δεδομένων δοκιμής.	Εάν η κατηγορική μεταβλητή έχει μια κατηγορία (στο σύνολο δεδομένων δοκιμής), η οποία δεν παρατηρήθηκε στο σύνολο δεδομένων εκπαίδευσης, τότε το μοντέλο θα αποδώσει πιθανότητα 0 (μηδέν) και δεν θα είναι σε θέση να κάνει πρόβλεψη. Αυτό είναι συχνά γνωστό ως "μηδενική συχνότητα". Για να το λύσουμε αυτό, μπορούμε να χρησιμοποιήσουμε την τεχνική εξομάλυνσης. Μια από τις απλούστερες τεχνικές εξομάλυνσης ονομάζεται εκτίμηση Laplace.
Αποδίδει επίσης καλά στην πρόβλεψη πολλαπλών κλάσεων.	Κακός εκτιμητής, οπότε οι έξοδοι πιθανότητας από το predict_proba δεν πρέπει να λαμβάνονται πολύ σοβαρά υπόψη.
Όταν ισχύει η υπόθεση της ανεξαρτησίας, ένας ταξινομητής Naive Bayes αποδίδει καλύτερα σε σύγκριση με άλλα μοντέλα όπως η λογιστική παλινδρόμηση και χρειάζονται λιγότερα δεδομένα εκπαίδευσης.	Η υπόθεση των ανεξάρτητων προβλεπτών. Στην πραγματική ζωή, είναι σχεδόν αδύνατο να έχουμε ένα σύνολο προβλεπτών που να είναι εντελώς ανεξάρτητοι.
Αποδίδει καλά στην περίπτωση κατηγορικών μεταβλητών εισόδου σε σύγκριση με αριθμητικές μεταβλητές. Για τις αριθμητικές μεταβλητές, υποτίθεται κανονική κατανομή (καμπύλη καμπάνας, η οποία είναι μια ισχυρή υπόθεση).	

5.1.4 Linear & Logistic Regression

Η λογιστική παλινδρόμηση [18] είναι ένας αλγόριθμος μάθησης με επίβλεψη που χρησιμοποιείται για την πρόβλεψη μιας εξαρτημένης κατηγορικής μεταβλητής-στόχου. Όταν υπάρχει ένα μεγάλο σύνολο δεδομένων που πρέπει να κατηγοριοποιηθεί, χρησιμοποιείται η λογιστική παλινδρόμηση.
Όταν έχουμε ένα παράδειγμα όπου υπάρχουν μόνο δύο πιθανές απαντήσεις, μιλάμε για δυαδική λογιστική παλινδρόμηση. Ωστόσο, είναι επίσης δυνατό να

ρυθμιστεί η λογιστική παλινδρόμηση με περισσότερες από δύο πιθανές κατηγορίες, οπότε τότε αναφερόμαστε στην πολυωνυμική λογιστική παλινδρόμηση.
Η λογιστική παλινδρόμηση απαιτεί, η εξαρτημένη μεταβλητή, να είναι κατηγορική. Το αποτέλεσμα είναι είτε Ναι είτε Οχι - δεν υπάρχει ενδιάμεσο εύρος. Η μονάδα μέτρησης διαφέρει από τη γραμμική παλινδρόμηση, καθώς παράγει μια πιθανότητα, αλλά η συνάρτηση logit μετατρέπει την καμπύλη S σε ευθεία γραμμή.
Ένα πρόβλημα που έχει συνεχές αποτέλεσμα, όπως η πρόβλεψη του βαθμού ενός μαθητή δεν είναι καλός υποψήφιος για τη χρήση της λογιστικής παλινδρόμησης. Άλλες επιλογές, όπως η γραμμική παλινδρόμηση μπορεί να είναι πιο κατάλληλες.
Τα μοντέλα γραμμικής παλινδρόμησης χρησιμοποιούνται για τον προσδιορισμό της σχέσης μεταξύ μιας συνεχούς εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Όταν υπάρχει μόνο μία ανεξάρτητη μεταβλητή και μία εξαρτημένη μεταβλητή, είναι γνωστή ως απλή γραμμική παλινδρόμηση, αλλά καθώς αυξάνεται ο αριθμός των ανεξάρτητων μεταβλητών, αναφέρεται ως πολλαπλή γραμμική παλινδρόμηση. Για κάθε τύπο γραμμικής παλινδρόμησης, επιδιώκεται η χάραξη μιας γραμμής καλύτερης προσαρμογής μέσω ενός συνόλου σημείων δεδομένων, η οποία συνήθως υπολογίζεται με τη μέθοδο των ελαχίστων τετραγώνων.
Ενώ και τα δύο μοντέλα χρησιμοποιούνται στην ανάλυση παλινδρόμησης για την πραγματοποίηση προβλέψεων σχετικά με μελλοντικά αποτελέσματα, η γραμμική παλινδρόμηση είναι συνήθως πιο κατανοητή. Η γραμμική παλινδρόμηση δεν απαιτεί επίσης τόσο μεγάλο μέγεθος δείγματος όσο η λογιστική παλινδρόμηση, χρειάζεται ένα επαρκές δείγμα για την αντιπροσώπευση τιμών σε όλες τις κατηγορίες απαντήσεων. Χωρίς ένα μεγαλύτερο, αντιπροσωπευτικό δείγμα, το μοντέλο μπορεί να μην έχει επαρκή στατιστική ισχύ για την ανίχνευση μιας σημαντικής επίδρασης.



Εικόνα 28: Linear and Logistic Regression [18]

Υπάρχουν τρεις τύποι μοντέλων λογιστικής παλινδρόμησης, με βάση την κατηγορική απόκριση:

Δυαδική λογιστική παλινδρόμηση (Binary logistic regression): Η εξαρτημένη μεταβλητή έχει μόνο δύο πιθανές εκβάσεις (π.χ. 0 ή 1). Στο πλαίσιο της λογιστικής παλινδρόμησης, αυτή είναι η πιο συχνά χρησιμοποιούμενη προσέγγιση, και γενικότερα, είναι ένας από τους πιο συνηθισμένους ταξινομητές για δυαδική ταξινόμηση.

Πολυωνυμική λογιστική παλινδρόμηση (Multinominal logistic regression): Η εξαρτημένη μεταβλητή έχει τρεις ή περισσότερες πιθανές εκβάσεις, χωρίς οι τιμές αυτές έχουν καθορισμένη σειρά.

Τακτική λογιστική παλινδρόμηση (Ordinal logistic regression): Χρησιμοποιείται όταν η μεταβλητή απόκρισης έχει τρία ή περισσότερα πιθανά αποτελέσματα, αλλά στην περίπτωση αυτή, οι τιμές αυτές έχουν καθορισμένη σειρά.

18 https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Πίνακας 17: Linear & Logistic Regression Pros and Cons

Πλεονεκτήματα	Μειονεκτήματα
Η λογιστική παλινδρόμηση είναι ευκολότερη στην εφαρμογή, την ερμηνεία και πολύ αποτελεσματική στην εκπαίδευση. Εάν ο αριθμός των παρατηρήσεων είναι μικρότερος από τον αριθμό των χαρακτηριστικών, η λογιστική παλινδρόμηση δεν πρέπει να χρησιμοποιείται, διαφορετικά, μπορεί να οδηγήσει σε υπερπροσαρμογή.	Δεν κάνει υποθέσεις σχετικά με τις κατανομές των κλάσεων στο χώρο των χαρακτηριστικών. Κατασκευάζει γραμμικά όρια.
Μπορεί εύκολα να επεκταθεί σε πολλαπλές κλάσεις (πολυωνυμική παλινδρόμηση) και μια φυσική πιθανολογική θεώρηση των προβλέψεων κλάσεων.	Ο σημαντικότερος περιορισμός της λογιστικής παλινδρόμησης είναι η υπόθεση της γραμμικότητας μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών.
Δεν παρέχει μόνο ένα μέτρο του πόσο κατάλληλος είναι ένας προγνωστικός παράγοντας (μέγεθος συντελεστή) αλλά και την κατεύθυνση της συσχέτισής του (θετική ή αρνητική).	Μπορεί να χρησιμοποιηθεί μόνο για την πρόβλεψη διακριτών συναρτήσεων. Ως εκ τούτου, η εξαρτημένη μεταβλητή της λογιστικής παλινδρόμησης δεσμεύεται στο σύνολο των διακριτών αριθμών.
Είναι πολύ γρήγορη στην ταξινόμηση άγνωστων δεδομένων.	Τα μη γραμμικά προβλήματα δεν μπορούν να επιλυθούν με τη λογιστική παλινδρόμηση επειδή έχει γραμμική επιφάνεια απόφασης. Γραμμικά διαχωρίσιμα δεδομένα σπάνια συναντώνται σε σενάρια του πραγματικού κόσμου.
Έχει ακρίβεια για πολλά απλά σύνολα δεδομένων και αποδίδει καλά όταν το σύνολο δεδομένων είναι γραμμικά διαχωρίσιμο.	Η λογιστική παλινδρόμηση απαιτεί μέση ή καθόλου πολυσυγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών.

Η λογιστική παλινδρόμηση είναι λιγότερο επιρρεπής στην υπερπροσαρμογή, αλλά μπορεί να υπερπροσαρμοστεί σε σύνολα δεδομένων υψηλής διάστασης.	Στη γραμμική παλινδρόμηση οι ανεξάρτητες και οι εξαρτημένες μεταβλητές συνδέονται γραμμικά. Αλλά η λογιστική παλινδρόμηση απαιτεί οι ανεξάρτητες μεταβλητές να σχετίζονται γραμμικά με τις λογαριθμικές αποδόσεις (log(p/(1-p)).

5.1.5 SVM

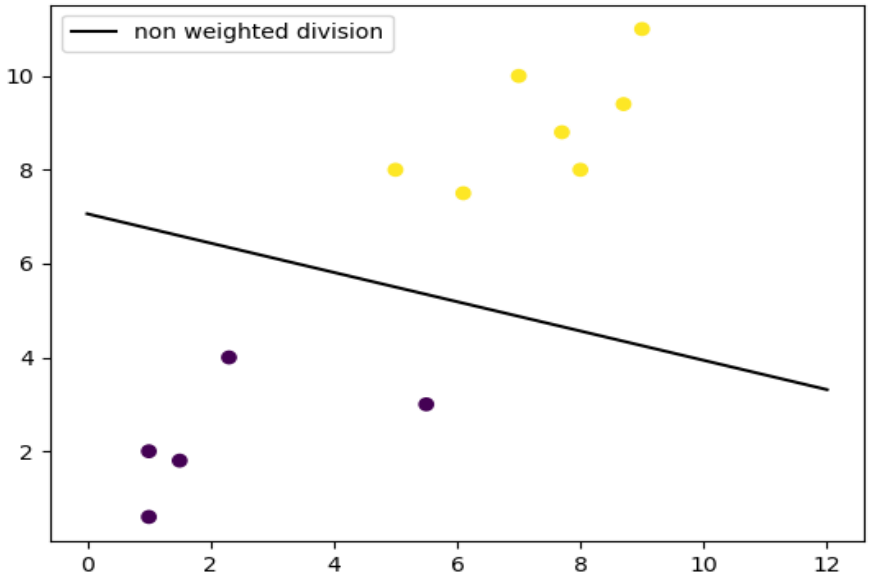
Ο SVM [19] είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη, είναι μια αναπαράσταση των δεδομένων ως σημεία στο χώρο, που απεικονίζονται έτσι ώστε τα δεδομένα των ξεχωριστών κατηγοριών να χωρίζονται από ένα σαφές κενό που είναι όσο το δυνατόν ευρύτερο. Στη συνέχεια, τα νέα δεδομένα απεικονίζονται στον ίδιο χώρο και θα ανήκουν σε μια κατηγορία με βάση το σε ποια πλευρά του κενού εμπίπτουν.

Εκτός από την εκτέλεση γραμμικής ταξινόμησης, οι SVM μπορούν να εκτελέσουν αποτελεσματικά και μη γραμμική ταξινόμηση, χρησιμοποιώντας το τέχνασμα πυρήνα, απεικονίζοντας τις εισόδους τους σε χώρους χαρακτηριστικών υψηλής διάστασης. Τα μη γραμμικά δεδομένα είναι βασικά τα οποία δεν μπορούν να διαχωριστούν με μια ευθεία γραμμή.

Υπάρχουν δύο διαφορετικοί τύποι SVM, που χρησιμοποιούνται:

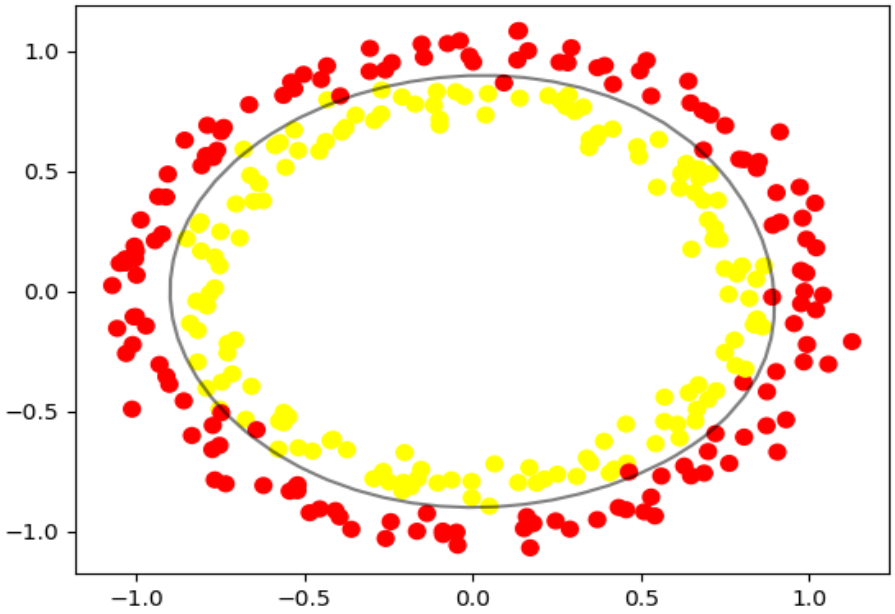
1. SVM: Απλό SVM: Συνήθως χρησιμοποιείται για προβλήματα γραμμικής παλινδρόμησης και ταξινόμησης.

2. Kernel SVM: Έχει μεγαλύτερη ευελιξία για μη γραμμικά δεδομένα επειδή μπορείτε να προσθέσετε περισσότερα χαρακτηριστικά για να προσαρμόσετε ένα υπερεπίπεδο αντί για ένα δισδιάστατο χώρο.



Εικόνα 29: Γραμμικη SVM [19]

19 https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/



Εικόνα 30: Μη γραμμική SVM με RBF kernel [20]

Πίνακας 18: SVM Pros and Cons

Πλεονεκτηματα:	**Μειονεκτήματα:**
Είναι χρήσιμο τόσο για γραμμικά διαχωρίσιμα (hard margin) όσο και για μη γραμμικά διαχωρίσιμα (soft margin) δεδομένα.	Η επιλογή του σωστού πυρήνα και των παραμέτρων μπορεί να είναι υπολογιστικά εντατική.
Είναι αποτελεσματική σε χώρους υψηλής διάστασης.	Δεν αποδίδει πολύ καλά, όταν στο σύνολο δεδομένων έχει κλάσεις-στόχους που επικαλύπτονται.
Είναι αποτελεσματική σε περιπτώσεις όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων.	Δεν παρέχει άμεσα εκτιμήσεις πιθανοτήτων, αυτές υπολογίζονται με τη χρήση μιας πενταπλής διασταυρούμενης επικύρωσης.

20 https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

Χρησιμοποιεί ένα υποσύνολο των σημείων εκπαίδευσης στη συνάρτηση απόφασης (που ονομάζονται διανύσματα υποστήριξης), οπότε είναι αποδοτική στη μνήμη.	

5.2 Εφαρμογή αλγορίθμων ML

Στο παράδειγμά μας η στήλη disaster_Size είχε καταχωρήσεις που ήταν μορφής string. Αντιστοιχήθηκαν στις τιμές της, float μεταβλητές με τον εξής τρόπο: large = 0, medium = 1, small = 2 .
Για αν γίνει αυτό πρέπει να γίνει import η skelarn, μία βιβλιοθήκη που περιέχει εργαλεία πρόβλεψης και ανάλυσης δεδομένων, όπως το sklearn.tree που χρησιμοποιήθηκε εδώ για το δέντρο αποφάσεων και το preprocessing το οποίο προετοιμάζει τα δεδομένα για την επεξεργασία που θα ακολουθηθεί στα επόμενα βήματα.
Ο αλγόριθμος πρέπει να εκπαιδεύσει ένα σύστημα το οποίο δέχεται ως input τα δεδομένα από τις στήλες casualties και injuredpeoplenumber δηλαδή τα θύματα και τους τραυματίες που προήλθαν από την συγκεκριμένη καταστροφή και να εξάγει ως output disaster_Size το οποίο δείχνει την τάξη μεγέθους της καταστροφής.

5.3 Αποτελέσματα Αλγορίθμων Μηχανικής Μάθησης

Δοκιμάστηκαν και οι πέντε αλγόριθμοι ταξινόμησης και παρακάτω παρουσιάζονται τα αποτελέσματά τους:

```
In [22]: from sklearn.metrics import classification_report, confusion_matrix
         print(confusion_matrix(y_test,y_pred))
         print(classification_report(y_test,y_pred))

[[   6  186    0]
 [   0 1311    0]
 [   0  704    0]]
              precision    recall  f1-score   support

           0       1.00      0.03      0.06       192
           1       0.60      1.00      0.75      1311
           2       0.00      0.00      0.00       704

    accuracy                           0.60      2207
   macro avg       0.53      0.34      0.27      2207
weighted avg       0.44      0.60      0.45      2207
```

Εικόνα 31: SVM

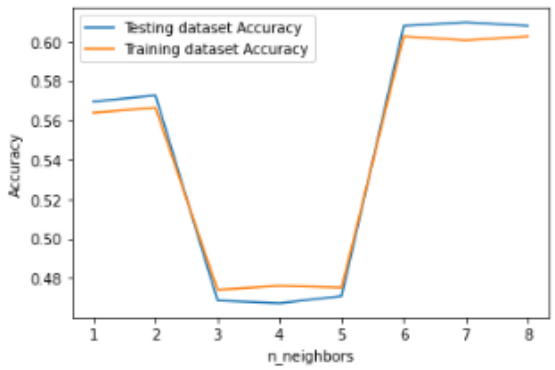
```
In [100]: model.fit(X_train, y_train)
Out[100]: DecisionTreeClassifier()

In [101]: preds = model.predict(X_test)

In [104]: np.unique(preds)
Out[104]: array([0, 1, 2])

In [105]: from sklearn.metrics import accuracy_score

In [106]: accuracy_score(y_test, preds)
Out[106]: 0.6080652469415496
```

Εικόνα 32: Decision Tree

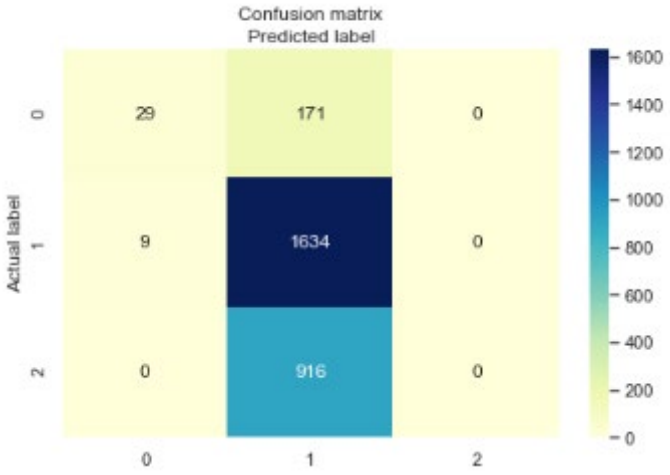
	precision	recall	f1-score	support
0	0.59	0.21	0.31	263
1	0.73	0.15	0.25	1949
2	0.37	0.96	0.54	1098
accuracy			0.42	3310
macro avg	0.56	0.44	0.37	3310
weighted avg	0.60	0.42	0.35	3310

Εικόνα 33: Naïve Bayes Algorithm

```
neighbors = np.arange(1, 9)
train_accuracy = np.empty(len(neighbors))
test_accuracy = np.empty(len(neighbors))

for i, k in enumerate(neighbors):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    train_accuracy[i] = knn.score(X_train, y_train)
    test_accuracy[i] = knn.score(X_test, y_test)
plt.plot(neighbors, test_accuracy, label = 'Testing dataset Accuracy')
plt.plot(neighbors, train_accuracy, label = 'Training dataset Accuracy')

plt.legend()
plt.xlabel('n_neighbors')
plt.ylabel('Accuracy')
plt.show()
```



Εικόνα 34: K-Nearest neighbours



Εικόνα 35:Logistic Regression

5.4 Ανάλυση Αποτελεσμάτων αλγορίθμων Μηχανικής μάθησης

Παρατηρήθηκε ότι οι προβλέψεις δεν μπορούν να θεωρηθούν ως ιδανικές, καθώς τα ποσοστά πρόβλεψης ήταν πολύ χαμηλά. Προκειμένου να βελτιωθεί το μοντέλο πρόβλεψης, εφαρμοστήκαν διάφορες υπερπαράμετροι [20] (hyperparameters), αλλά τα αποτελέσματα δεν βελτιώθηκαν σημαντικά. Έτσι, κρίθηκε ότι έπρεπε να δοκιμαστούν οι αλγόριθμοι σε μία άλλη βάση, καθώς τα δεδομένα φάνηκαν ανεπαρκή και ελλιπή.
Επιλέχθηκε ως κατάλληλη η βάση FireservceGR, και εφαρμόσθηκε το μοντέλο πρόβλεψης στη στήλη disasterSize. Δόθηκαν ως input στο μοντέλο οι στήλες carForce, manpowerForce, injuredpeoplenumber και casualties. Στόχος ήταν να αναπτυχθεί ένα μοντέλο, το οποίο να προβλέπει το μέγεθος μίας καταστροφής με βάση τις δυνάμεις που στάλθηκαν, τα οχήματά τους, τους τραυματίες και τα θύματα αυτής. Στη στήλη disasterSize έγιναν κάποιες παραδοχές που αφορούσαν τις καταχωρήσεις της. Για παράδειγμα, οι κλήσεις για απεγκλωβισμό από ανελκυστήρα κατηγοριοποιήθηκαν ως μικρές και οι κλήσεις για την ανεύρεση ενός αγνοούμενου ανθρώπου, ως μεγάλες.
Έμειναν έτσι τρεις κατηγορίες: μικρή, μεσαία, μεγάλη. Έπειτα, αντιστοιχήθηκαν σε αριθμητικά δεδομένα ως 2, 1 και 0 αντίστοιχα. Τα αποτελέσματα της εφαρμογής του κάθε αλγορίθμου παρουσιάζονται στις παρακάτω εικόνες.

	precision	recall	f1-score	support
0	0.75	0.25	0.37	53993
1	0.18	0.00	0.00	53541
2	0.37	0.98	0.54	53749
accuracy			0.41	161283
macro avg	0.43	0.41	0.30	161283
weighted avg	0.43	0.41	0.30	161283

Εικόνα 36: Naïve Bayes FireserviceGR

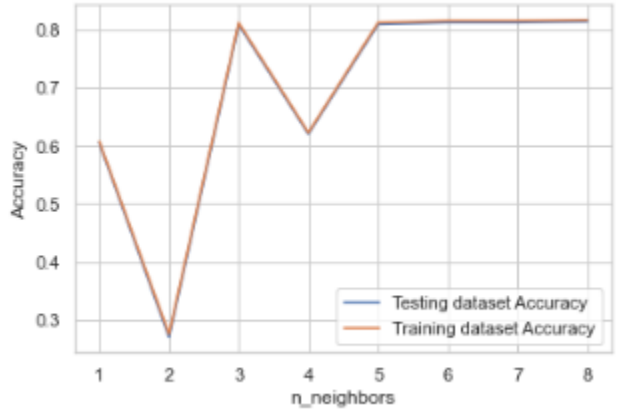
	precision	recall	f1-score	support
0	0.56	0.57	0.56	53993
1	0.00	0.00	0.00	53541
2	0.40	0.78	0.53	53749
accuracy			0.45	161283
macro avg	0.32	0.45	0.36	161283
weighted avg	0.32	0.45	0.37	161283

Εικόνα 37: Logistic Regression FireserviceGR

	precision	recall	f1-score	support
0	0.58	0.55	0.56	53993
1	0.43	0.48	0.45	53541
2	0.43	0.41	0.42	53749
accuracy			0.48	161283
macro avg	0.48	0.48	0.48	161283
weighted avg	0.48	0.48	0.48	161283

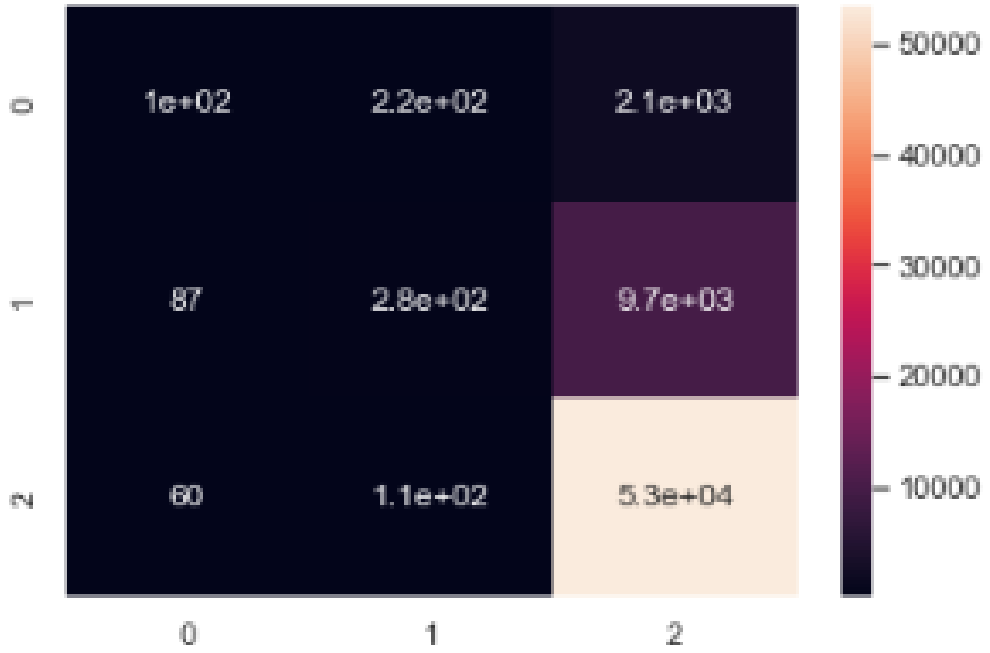
Εικόνα 38: DecisionTree FireserviceGR

```
import matplotlib.pyplot as plt
neighbors = np.arange(1, 9)
train_accuracy = np.empty(len(neighbors))
test_accuracy = np.empty(len(neighbors))

for i, k in enumerate(neighbors):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    train_accuracy[i] = knn.score(X_train, y_train)
    test_accuracy[i] = knn.score(X_test, y_test)
plt.plot(neighbors, test_accuracy, label = 'Testing dataset Accuracy')
plt.plot(neighbors, train_accuracy, label = 'Training dataset Accuracy')

plt.legend()
plt.xlabel('n_neighbors')
plt.ylabel('Accuracy')
plt.show()
```



Εικόνα 39: K-Nearest Neighbours FireserviceGR

```
[[  101   220  2128]
 [   87   278  9744]
 [   60   114 53480]]
```



Εικόνα 40: K-Nearest Neighbours FireserviceGR

0	0.60	0.02	0.04	145
1	0.60	1.00	0.75	1316
2	0.00	0.00	0.00	746
accuracy			0.60	2207
macro avg	0.40	0.34	0.26	2207
weighted avg	0.40	0.60	0.45	2207

Εικόνα 41: SVM FireserviceGR

Σε σχέση με την προηγούμενη βάση (Global Landslide Catalog), δόθηκαν παραπάνω στήλες ως input και παρατηρήθηκε ότι τα αποτελέσματα ήταν αρκετά καλύτερα και πιο αξιόπιστα. Για να βελτιστοποιηθεί η διαδικασία ακολουθήθηκε η χρήση της μεθόδου resampling[21]. Η μέθοδος αυτή χρησιμοποιεί διάφορες τεχνικές για τη συλλογή περισσότερων πληροφοριών σχετικά με ένα συγκεκριμένο δείγμα. Δύο τέτοιες περιπτώσεις είναι η εκ νέου λήψη του δείγματος ή η εκτίμηση της ακρίβειάς του όπου X και y είναι τα features και τα labels αντίστοιχα. Εφαρμόζουμε τη μέθοδο resampling (oversampling συγκεκριμένα) στις minority κλασεις (που εμφανίζονται λιγότερες φορές) ώστε να φτάσουν τα παραδείγματα τους, μαζι με τα labels φυσικά, τη majority class. Στο συγκεκριμένο παράδειγμα έχουμε πολύ περισσότερα παραδείγματα της κλάσης 2 (Μικρό μέγεθος καταστροφής) και θέλουμε να δημιουργήσουμε αντίστοιχο αριθμό παραδειγμάτων και στις κλάσεις 0 και 1. Δημιουργούνται παραδείγματα του τύπου 5 τραυματίες, 3 νεκροί, 5 οχήματα που είναι ένα "νέο" X και Μεσαία για αυτό το παράδειγμα που είναι ένα νέο y που αντιστοιχεί σε αυτό το X. (συνήθως το κάνει με επανατοποθέτηση των παραδειγμάτων των minority κλάσεων).

Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

6 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΒΕΛΤΙΩΣΕΙΣ

Έγινε εκτενής αναζήτηση βάσεων που περιείχαν δεδομένα καταστροφής σε ιδιωτικές εταιρείες αλλά και σε κρατικούς φορείς. Παρατηρήθηκε ότι οι κρατικοί φορείς παρείχαν τις βάσεις χωρίς να απαιτείται συνδρομή από τον χρήστη ενώ αυτό δεν ίσχυε για όλες τις ιδιωτικές εταιρείες. Σε πολλές περιπτώσεις δινόταν πρόσβαση σε ένα συγκεκριμένο τμήμα των πληροφοριών, προκειμένου ο χρήστης να αγοράσει την προνομιούχα συνδρομή. Έτσι, οι εννέα βάσεις που αρχικά επιλέχτηκαν έπρεπε να έχουν ελεύθερη πρόσβαση για την έρευνα που πραγματοποιήθηκε αλλά και για μελλοντική χρήση αυτών.

Οι εννέα βάσεις μελετήθηκαν και εντοπίστηκαν διάφορα προβλήματα σε αυτές. Ένα πρόβλημα που παρατηρήθηκε σε αρκετές εξ αυτών ήταν η ύπαρξη περιττών δεδομένων και κενών στηλών. Επίσης, κάποιες βάσεις περιείχαν δεδομένα που αφορούσαν ένα πολύ μικρό χρονικό διάστημα, γεγονός που τις καθιστούσε αναξιόπιστες για την εξαγωγή γενικότερων συμπερασμάτων. Ένα άλλο πρόβλημα αποτέλεσε το γεγονός ότι η ημερομηνία του συμβάντος δινόταν σε μη επιθυμητή μορφή ή δεν υπήρχε καθόλου ως καταχώρηση.

Προκειμένου να επιλυθούν τα ζητήματα αυτά, ακολουθήθηκαν συγκεκριμένα κοινά βήματα, προσαρμοσμένα στην εκάστοτε βάση. Αρχικά, αφαιρέθηκαν οι στήλες με τα περιττά δεδομένα και τις μηδενικές καταχωρήσεις. Εφ’ όσον κάποια βάση κρίθηκε ελλιπής δεν επιλέχθηκε για περαιτέρω ανάλυση καθώς δεν γινόταν να προστεθούν δεδομένα σε αυτήν. Για την ακριβή ημερομηνία του συμβάντος, τροποποιώντας τα δεδομένα που δόθηκαν σε μη επιθυμητή μορφή, δημιουργήθηκαν στήλες που περιείχαν την ημέρα, το μήνα και τη χρονολογία έναρξης και λήξης αντίστοιχα. Έτσι επιλέχθηκαν πέντε βάσεις που πληρούσαν τα κριτήρια που τέθηκαν και μπορούσαν να εφαρμοστούν οι τροποποιήσεις σε αυτές.

Στη συνέχεια, οι έξι βάσεις έπρεπε να ακολουθήσουν ένα συγκεκριμένο πρότυπο που αφορά τα γνωρίσματά τους, προκειμένου στο τέλος να περιέχουν ίδια πεδία και ίδιο τύπο δεδομένων. Σε περίπτωση που τα ονόματα για τα ίδια πεδία διέφεραν, μετονομάστηκαν με χρήση εντολών pandas. Επίσης, άλλαξε ο τύπος δεδομένων ανάλογα τα περιεχόμενα της στήλης. Για παράδειγμα, οι ανθρώπινες απώλειες και οι τραυματίες ήταν τύπου int64, ενώ οι ημερομηνίες έναρξης και λήξης, τύπου object. Παρατηρήθηκε ότι ενώ πολλές βάσεις φαινομενικά περιείχαν ίδια δεδομένα, έπρεπε να αλλάξει ο τύπος τους γιατί δεν μπορούσε να γίνει αντιστοίχιση αυτών με τις υπόλοιπες.

Με χρήση αλγορίθμων μηχανικής μάθησης, έπρεπε να δημιουργηθεί ένα μοντέλο πρόβλεψης. Επιλέχθηκαν οι 5 αλγόριθμοι ταξινόμησης που δοκιμάστηκαν οι οποίοι είναι οι: SVM, Naïve Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbour. Αυτοί εφαρμοστήκαν στη στήλη Disaster-Size της βάσης Global Landslide Catalog, προκειμένου να δημιουργηθεί το επιθυμητό μοντέλο πρόβλεψης. Η συγκεκριμένη στήλη περιείχε δεδομένα τύπου object τα οποία μετατράπηκαν σε int. Για να λειτουργήσει το μοντέλο, βασίστηκε στις στήλες casualties και injuredpeoplenumber. Δηλαδή, βάσει των στοιχείων των δύο στηλών αυτών έβγαινε μια πρόβλεψη για την καταχώρηση της στήλης Disaster_Size με πιθανότητα επιτυχίας η οποία κυμαινόταν από 40-

65% ανάλογα με τον αλγόριθμο που εφαρμόστηκε. Τα αποτελέσματα δεν ήταν τα αναμενόμενα, αν και φαινομενικά η μεθοδολογία ήταν σωστή, οπότε εφαρμόστηκε η ίδια τεχνική σε άλλα δεδομένα.
Η δεύτερη βάση που επιχειρήθηκε εφαρμογή των αλγορίθμων ήταν η FireserviceGR. Λόγω ύπαρξης περισσότερων στηλών ως input, έγινε πρόβλεψη με μεγαλύτερη ακρίβεια της στήλης disatserSize. Όταν η παρατήρηση σε μία κλάση είναι υψηλότερη από την παρατήρηση στις άλλες κλάσεις, τότε υπάρχει το φαινόμενο class imbalance. Στο συγκεκριμένο παράδειγμα υπήρχαν χιλιάδες περισσότερες καταχωρήσεις με την τιμή 'Μικρή' ή 2 μετά την μετατροπή που έγινε. Ακόμα και μετά την εφαρμογή υπερπαραμέτρων και της μεθόδου resampling τα αποτελέσματα, αν και ξεπερνούσαν τη τυχαιότητα που θα είχε 33,333% σε περίπτωση μη εφαρμογής μοντέλου προβλέψεων, δεν θεωρήθηκαν ικανοποιητικά. Αυτό είναι το φαινόμενο garbage in, garbage out, που δεν αφορά τη λειτουργία του αλγορίθμου, αλλά τα δεδομένα που εισάγονται και κατ' επέκταση, που εξάγονται.
Έτσι, καταλήγουμε στο συμπέρασμα ότι υπάρχει περιθώριο για μελλοντικές βελτιώσεις. Μία βελτίωση, είναι τα δεδομένα να έχουν πληθώρα καταχωρήσεων και να χρησιμοποιούνται πολλές πλήρεις στήλες ως input. Επίσης, οι στήλες αυτές να μην έχουν πολλές κενές καταχωρήσεις ή περιττά δεδομένα, τα οποία ακόμα και να εξαλειφτούν δημιουργούν προβλήματα.
Η χρήση υπερπαραμέτρων (hyperparameter tuning), όπως έγινε στη συγκεκριμένη εφαρμογή, βελτιώνει την έκβαση του αποτελέσματος του μοντέλου. Μια μελλοντική βελτίωση είναι η εφαρμογή και άλλων υπερπαραμέτρων στον αντίστοιχο αλγόριθμο ταξινόμησης προσαρμοσμένων στα κριτήρια επιλογής των input στηλών disaster_Size GLC kai FireServiceGR.
Για την μεγαλύτερη αυτοματοποίηση της διαδικασίας είναι δυνατή η χρήση docker to application[22] προκειμένου να εκτελείται αυτόματα μέσω docker η κλήση της βάσης και η επεξεργασία αυτής μέσω αλγορίθμων μηχανικής μάθησης, επισπεύδοντας την διαδικασία και διευκολύνοντας έτσι τον χρήστη.
Συμπερασματικά, αν και τα αποτελέσματα του μοντέλου πρόβλεψης δεν ήταν άρτια και επιδέχονται τις τροποποιήσεις που αναφέρθηκαν, ήταν αρκετά πάνω από το baseline. Έτσι, μπορούν να εξαχθούν τα επιθυμητά αποτελέσματα από την εφαρμογή του και σε άλλες βάσεις με αντίστοιχα κριτήρια επιλογής και τροποποίησης αυτών.

Ο πλήρης κώδικας που χρησιμοποιήθηκε για την διπλωματική εργασία υπάρχει διαθέσιμος στο:
https://github.com/Vigsanir/DisasterDatabases

Σχεδιασμός συστήματος συλλογής δεδομένων, κατασκευή προγραμματιστικών διεπαφών και
προβλεπτικών μεθόδων για δεδομένα καταστροφών

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] 'Welcome | EM-DAT'. https://www.emdat.be/welcome (accessed Oct. 23, 2022).

[2] G. Palm, 'Warren McCulloch and Walter Pitts: A Logical Calculus of the Ideas Immanent in Nervous Activity', *Brain Theory*, pp. 229–230, 1986, doi: 10.1007/978-3-642-70911-1_14.

[3] I. Kar-Purkayastha, M. Clarke, and V. Murray, 'Dealing with disaster databases – What can we learn from health and systematic reviews?', *PLOS Currents Disasters*, 2011, doi: 10.1371/CURRENTS.RRN1272.

[4] 'World Health Organization (WHO)'. https://www.who.int/ (accessed Oct. 21, 2022).

[5] C. Huggel, A. Raissig, M. Rohrer, G. Romero, A. Diaz, and N. Salzmann, 'How useful and reliable are disaster databases in the context of climate and global change? A comparative case study analysis in Peru', *Natural Hazards and Earth System Sciences*, vol. 15, no. 3, pp. 475–485, Mar. 2015, doi: 10.5194/NHESS-15-475-2015.

[6] N. Smith, 'Research Guides: Natural Disasters: A Resource Guide: Databases and Journal Articles', Accessed: Oct. 22, 2022. [Online]. Available: https://guides.loc.gov/natural-disasters/articles-databases

[7] 'Global Internal Displacement Database | IDMC'. https://www.internal-displacement.org/database/displacement-data (accessed Oct. 15, 2022).

[8] D. Kirschbaum, T. Stanley, and Y. Zhou, 'Spatial and temporal analysis of a global landslide catalog', *Geomorphology*, vol. 249, pp. 4–15, Nov. 2015, doi: 10.1016/J.GEOMORPH.2015.03.016.

[9] M. Meghraoui *et al.*, 'Evidence for 830 years of seismic quiescence from palaeoseismology, archaeoseismology and historical seismicity along the Dead Sea fault in Syria', *Earth Planet Sci Lett*, vol. 210, no. 1–2, pp. 35–52, May 2003, doi: 10.1016/S0012-821X(03)00144-4.

[10] '1. What is the Jupyter Notebook? — Jupyter/IPython Notebook Quick Start Guide 0.1 documentation'. https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html (accessed Oct. 21, 2022).

[11] 'What is NumPy? — NumPy v1.23 Manual'. https://numpy.org/doc/stable/user/whatisnumpy.html (accessed Oct. 21, 2022).

[12] 'pandas · PyPI'. https://pypi.org/project/pandas/ (accessed Oct. 21, 2022).

[13] 'FastAPI'. https://fastapi.tiangolo.com/ (accessed Oct. 21, 2022).

[14] 'Machine Learning Algorithms - Giuseppe Bonaccorso - Google Books'. https://books.google.gr/books?hl=en&lr=&id=_-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+algorithms&ots=epjDF2BDaK&sig=IS4g68K8H_U5jKA07eFQ1eezukE&redir_esc=y#v=onepage&q=machine%20learning%20algorithms&f=false (accessed Oct. 15, 2022).
[15] Z.-H. Zhou, 'Decision Trees', *Mach Learn*, pp. 79–102, 2021, doi: 10.1007/978-981-15-1967-3_4.
[16] 'What is the k-nearest neighbors algorithm? | IBM'. https://www.ibm.com/topics/knn (accessed Oct. 21, 2022).
[17] 'Naïve Bayes Algorithm: Everything You Need to Know - KDnuggets'. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html (accessed Oct. 17, 2022).
[18] 'How Does Linear And Logistic Regression Work In Machine Learning? | Analytics Steps'. https://www.analyticssteps.com/blogs/how-does-linear-and-logistic-regression-work-machine-learning (accessed Oct. 21, 2022).
[19] 'Support Vector Machine (SVM) Algorithm - Javatpoint'. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm (accessed Oct. 21, 2022).
[20] 'Hyperparameter Optimization for Machine Learning Models - KDnuggets'. https://www.kdnuggets.com/2020/05/hyperparameter-optimization-machine-learning-models.html (accessed Oct. 22, 2022).
[21] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, 'A Resampling Method for Imbalanced Datasets Considering Noise and Overlap', *Procedia Comput Sci*, vol. 176, pp. 420–429, Jan. 2020, doi: 10.1016/J.PROCS.2020.08.043.
[22] 'Containerize an application | Docker Documentation'. https://docs.docker.com/get-started/02_our_app/ (accessed Oct. 23, 2022).