ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ

<u>ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ</u>

ΧΑΡΑΥΓΗΣ ΑΛΕΞΑΝΔΡΟΣ - *713242017067*

ΒΟΛΙΩΤΗΣ ΙΩΑΝΝΗΣ ΑΛΕΞΑΝΔΡΟΣ - *171062*

ΕΝΡΙΚ ΜΑΖΑΙ - *18390080*

1. ***ΕΙΣΑΓΩΓΗ***

Έχοντας διεξάγει μια εκτενή έρευνα και εξετάσει διάφορες διαθέσιμες ιστοσελίδες με δεδομένα μεγάλης κλίμακας για ανάλυση, καταλήξαμε στο συμπέρασμα ότι το φύλλο δεδομένων με τίτλο "Data Science Salaries 2023" είναι το πιο κατάλληλο και ενδιαφέρον για την ανάλυσή μας. Αυτό το σύνολο δεδομένων αναφέρεται στις μισθολογικές τάσεις των υπαλλήλων στον τομέα της επιστήμης των δεδομένων.

Χρησιμοποιώντας τη γλώσσα προγραμματισμού Python και διάφορες βιβλιοθήκες, πραγματοποιούμε την εξόρυξη και ανάλυση των δεδομένων. Εκτελούμε διάφορες τεχνικές ανάλυσης όπως απεικόνιση γραφημάτων και γραφικών παραστάσεων, εντοπίζουμε συσχετίσεις μεταξύ των μεταβλητών και εκτελούμε προβλέψεις με χρήση αλγορίθμων ταξινόμησης και παλινδρόμησης.

2. ***ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΚΙΝΗΤΡΟ***

Τα κίνητρα που μας οδήγησαν στην επιλογή του συγκεκριμένου θέματος "DATA SCIENCE SALARIES 2023" για την εργασία μας είναι αρκετά. Μεταξύ άλλων ορισμένοι από τους βασικούς στους λόγους επιλογής περιλαμβάνονται και οι εξής:

1. Υπάρχει γενικότερο ενδιαφέρον και θέληση για ανακάλυψη της επιστήμης των δεδομένων τόσο από το μέρος των φοιτητών, των καθηγητών όσο και των επαγγελματιών που αποτελούν μέρος του συγκεκριμένου κλάδου.

2. Προοπτικές αγοράς εργασίας: Η επιστήμη των δεδομένων είναι μια αναπτυσσόμενη και ζωντανή βιομηχανία. Οι επαγγελματίες στον τομέα αυτόν έχουν συχνά καλές προοπτικές για καριέρα και μπορούν να αναζητήσουν ευκαιρίες για ανέλιξη και αύξηση του εισοδήματός τους.

3. Ανάγκη διερεύνησης των μισθολογικών τάσεων: Η κατανόηση των μισθολογικών τάσεων στον τομέα της επιστήμης των δεδομένων μπορεί να βοηθήσει επαγγελματίες ή φοιτητές να κατανοήσουν πόσο ανταγωνιστικός είναι ο τομέας και ποιες είναι οι περιοχές όπου μπορούν να αναπτύξουν τις δεξιότητές τους για να βελτιώσουν τις πιθανότητες για μεγαλύτερο εισόδημα.

4. Συνεισφορά στην κοινότητα: Η ανάλυση δεδομένων για τους μισθούς μπορεί να συμβάλει στην πληροφόρηση των άλλων επαγγελματιών ή ενδιαφερομένων για τις τάσεις της αγοράς και να παρέχει χρήσιμα δεδομένα για αποφάσεις στον τομέα της καριέρας.

Προφανέστατα, η ανάλυση δεδομένων των μισθολογικών τάσεων των υπαλλήλων στον τομέα της επιστήμης των δεδομένων είναι ακράδαντα χρήσιμη για τους εξής:

1 .Φοιτητές και εκπαιδευόμενους : Για όσους σπουδάζουν ή εκπαιδεύονται στον τομέα της επιστήμης των δεδομένων, η ανάλυση των μισθολογικών τάσεων μπορεί να παρέχει πληροφορίες σχετικά με τις προοπτικές καριέρας και την ανταγωνιστικότητα της αγοράς εργασίας. Μπορεί να τους βοηθήσει να κατανοήσουν ποιες δεξιότητες και γνώσεις είναι ιδιαίτερα ζητούμενες και ποια μπορεί να είναι η ανταμοιβή τους στην αγορά εργασίας.

2. Επαγγελματίες επιστήμονες δεδομένων: Οι επαγγελματίες της επιστήμης των δεδομένων μπορούν να χρησιμοποιήσουν την ανάλυση δεδομένων για τους μισθούς για να κατανοήσουν τις

τάσεις αμοιβής στον τομέα και να καταλάβουν ποιες δεξιότητες και εμπειρίες μπορούν να ενισχύσουν για να αυξήσουν την αξία τους στην αγορά εργασίας.

3. Εταιρείες και εργοδότες: Για τις εταιρείες που δραστηριοποιούνται στον τομέα της επιστήμης των δεδομένων, η ανάλυση των μισθολογικών δεδομένων μπορεί να βοηθήσει να κατανοήσουν τις τρέχουσες αγοραίες τάσεις και να προσελκύσουν και να διατηρήσουν ταλαντούχους επαγγελματίες. Επιπλέον, μπορεί να τους βοηθήσει στον σχεδιασμό πολιτικών αμοιβών και επιδοτήσεων για τους εργαζομένους τους.

4. Απλά ενδιαφερόμενοι: Άτομα που ενδιαφέρονται για τον γενικότερο τομέα της επιστήμης των δεδομένων μπορούν να βρουν ενδιαφέρουσες πληροφορίες σχετικά με τις αμοιβές, τις αγοραίες τάσεις και τις δυνατότητες καριέρας στον τομέα αυτόν μέσω της ανάλυσης δεδομένων για τους μισθούς

3. ***ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΣΑΜΕ***

Σχετικά με το σύνολο δεδομένων που χρησιμοποιούμε το dataset περιέχει 11 στήλες δεδομένων, η οποίες είναι:

work_year: Το έτος που καταβλήθηκε ο μισθός. (2021-2022-2023)

experience_level: Το επίπεδο εμπειρίας στην εργασία κατά τη διάρκεια του έτους. Περιέχει τις τιμές EN, MI, SE, EX που μεταφράζονται σε entry, intermediate ,senior ,executive αντίστοιχα.

job_type: Ο τύπος απασχόλησης κατά τη διάρκεια της έτους.Περιέχει κυρίως τιμές PT-FT για part-time και full-time.

job_title: Ο τίτλος εργασίας κατά τη διάρκεια του έτους.

salary: Ο συνολικός ακαθάριστος μισθός που καταβλήθηκε.

salary_currency: Το νόμισμα του μισθού που καταβάλλεται ως κωδικός νομίσματος.

salary_in_usd: Ο μισθός σε δολάρια

employee_residence: Η κύρια χώρα διαμονής του εργαζομένου ως κωδικός χώρας ISO 3166.

remote_ratio: Ο συνολικός όγκος της εργασίας που γίνεται εξ αποστάσεως. Οι τιμές είναι 100 για πλήρη εργασια εξ αποστάσεως , 50 για υβριδικό μοντέλο εργασίας και 0 για πλήρη on-site απασχόληση.

company_location: Η χώρα του κεντρικού γραφείου ή του υποκαταστήματος του εργοδότη.

company_size: Ο διάμεσος αριθμός ατόμων που εργάστηκαν για την εταιρεία κατά τη διάρκεια του έτους. Οι τιμές που συναντάμε είναι S,M,L για μικρό, μεσαία και μεγάλο μέγεθος αντίστοιχα.

4. ***ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΥ ΑΝΑΛΥΣΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ***

Η ανάλυση του dataset έγινε αρχικά με κάποια γραφήματα ώστε να καταλαβουμε τα δεδομένα μας καλύτερα και την σύνδεση των στηλών μεταξύ τους. Αυτά τα γραφήματα αργότερα θα μας βοηθήσουμε και στην καλύτερη προσέγγιση των συμπερασμάτων μας. Η στήλη που μας ενδιαφέρει περισσότερο είναι η στήλη με τους μισθούς των εργαζομένων (σε δολάρια) , όποτε εμφανίζουμε με μορφή boxplot τις σχέσεις των μισθών με το ποσοστό τηλεργασίας (remote ratio) και το μέγεθος της εταιρείας που εργάζονται (company size). Στην συνέχεια δείχνουμε τον μέσο μισθό στις χώρες που έχουμε πάνω από 15 εγγραφές και την σχέση του μισθού με το επίπεδο εμπειρίας του εργαζομένου (experience level).

Αφού έχουμε εκτελέσει τα απαραίτητα βήματα του καθαρισμού των δεδομένων μας σε περίπτωση κενών στοιχείων προχωράμε στο πείραμα μας. Θα χρησιμοποιήσουμε την τεχνική ανάλυσης classification & regression.

Για το **classification** χρησιμοποιούμε την στήλη salary_in_usd και την χωρίζουμε σε 3 classes με την βοήθεια τεταρτημορίων (0.25 , 0.75) όπου έχουμε : low (0 εώς 95.000 $) , medium (95.000-175.000$) , high (+ 175000$). Στην συνέχεια κρατάμε τις στήλες που μας ενδιαφέρουν και έχουν μεγαλύτερη σύνδεση με τον τελικό μισθό όπου είναι 'work_year', 'experience_level','employment_type', 'job_title','remote_ratio','company_location', 'company_size'. Με την βοήθεια του onehot encoder μπορούμε να αναγνωρίσουμε και τα μη αριθμητικά δεδομένα μετατρέποντας τα σε δυαδικές μορφές. Όταν ολοκληρωθεί και η προεπεξεργασία των κατηγορηματικών δεδομένων χωρίζουμε το επιλεγμένο dataset στο training και στο testing με ποστοστά 80% και 20% αντίστοιχα. Αυτό σημαίνει πως χρησιμοποιούμε το 80% του dataset για να εκπαιδεύσουμε το σύστημα μηχανικής εκμάθησης και έπειτα το αξιολογούμε με το 20% που δεν χρησιμοποιήθηκε. Η μέθοδος που προτιμήθηκε για την δημιουργία του classification model είναι η Logistic Regression λόγω καλύτερων αποτελεσμάτων.

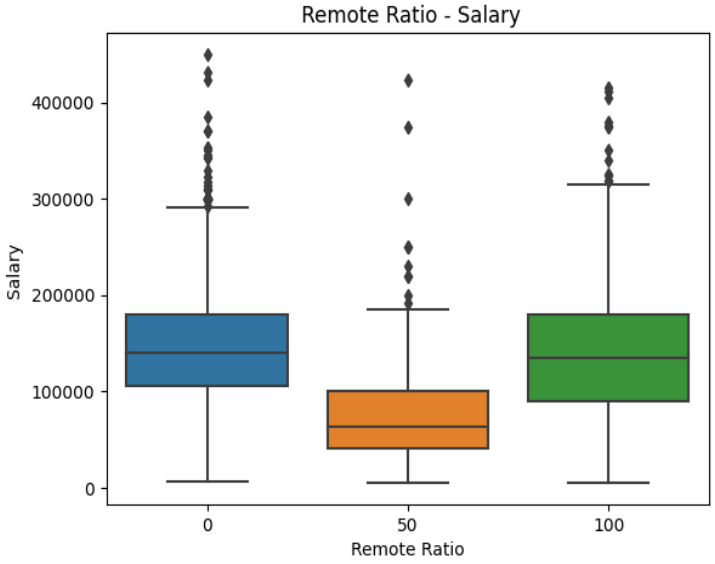
Για το **regression** ακολουθούμε παρόμοια βήματα παράλληλα με το classification στο κώδικα μας και χρησιμοποιούμε την μέθοδο Linear Regression για το μοντέλο μας. Χωρίζεται το dataset στα ίδια ποσοστά σε training - testing και χρησιμοποιούμε τις ίδιες στήλες με το classification model.

Ο σκοπός του classification και του Logistic Regression είναι να κατηγοριοποιήσει τα δεδομένα εισόδου προβλέποντας σε ποιά οικονομική κλάση ανήκουν. Ενώ ο σκοπος του regression και της Linear Regression είναι να προβλέψει σαν έξοδο τον ακριβή μισθό του εργαζομένου ανάλογα με τις τιμές εισόδου.

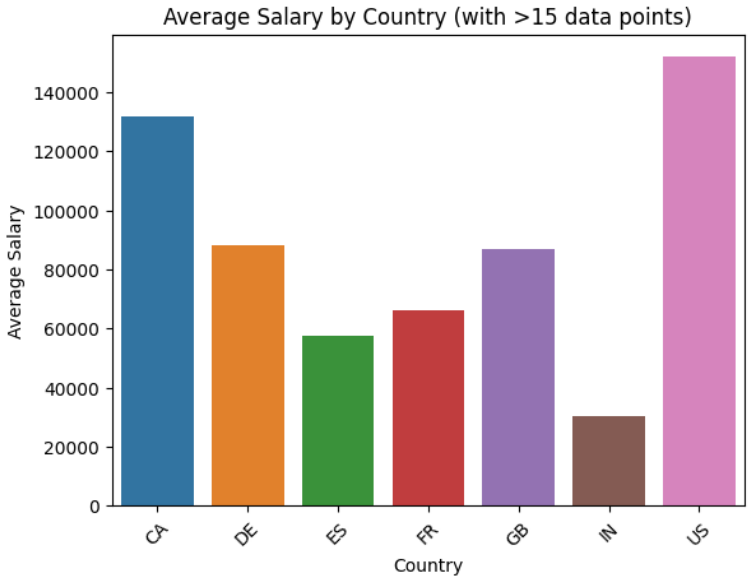
5. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικά απο την οπτικοποίηση των δεδομένων και τις σχέσει μεταξύ τους έχουμε τα παρακάτω γραφήματα που αναφέραμε

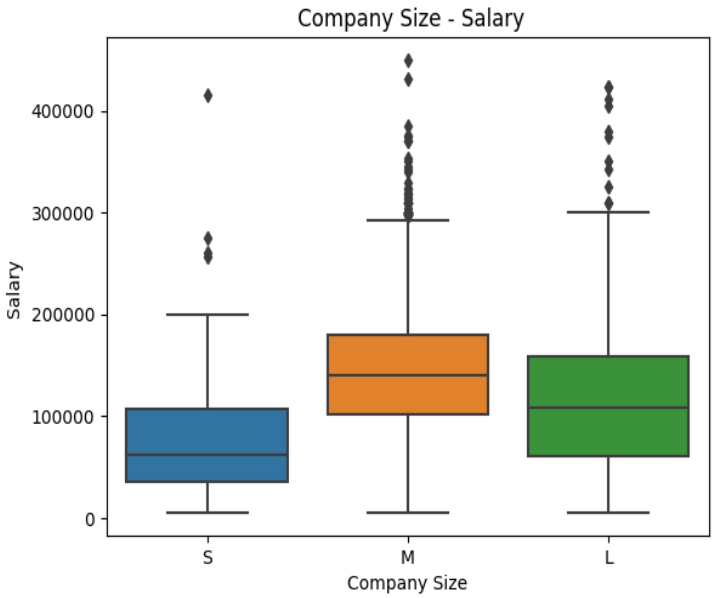
remote ratio-salary



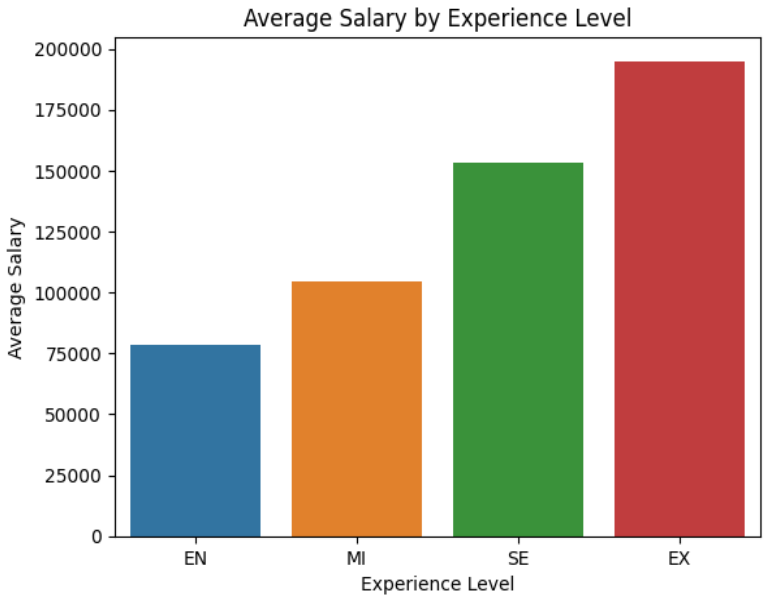
Μέσος μισθών χωρών με 15> εγγραφές



company size -salary



Μέσος μισθός βάση Εμπειρίας



Για τα αποτελέσματα των πειραματικών μοντέλων μας έχουμε δημιουργήσει στο κώδικα μια μεταβλητή new instance όπου του εισάγουμε τα δεδομένα που θέλουμε να έχει ο εργαζόμενος μας και με την έξοδο τα μοντέλα τον κατατάσουν σε μια κλάση (classification model) και υπολογίζουν τον μισθό του (regression model). Ακολουθούν κάποια παραδείγματα

1ο Παράδειγμα είσοδος-έξοδος

```
# Example instance
new_instance = [[2023,'EN', 'FT', 'ML Engineer',100,'US', 'M']]
```

```
Predicted Salary Class: ['medium'] Range: 95000.00-175000.00
Predicted Salary (in number): 148048.92
```

2ο Παράδειγμα

```
# Example instance
new_instance = [[2023,'EX', 'FT', 'ML Engineer',100,'US', 'M']]
```

```
Predicted Salary Class: ['high'] Range: 175000.00+
Predicted Salary (in number): 235727.69
```

3ο Παράδειγμα

```
# Example instance
new_instance = [[2023,'MI', 'FT', 'ML Engineer',100,'IN', 'M']]
```

```
Predicted Salary Class: ['low'] Range: 0-95000.00
Predicted Salary (in number): 82726.00
```

6. ***ΚΡΙΤΙΚΗ ΑΠΟΤΙΜΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ***

Με μια πρώτη ματιά στις εξόδους των μοντέλων μπορούμε να πούμε πως ανταποκρίνεται σωστά στις εισόδους μας. Αναλυτικότερα κοιτώντας τα γραφήματα που δημιουργήσαμε για τις σχέσεις των στηλών με τον μισθών μπορούμε να δούμε πως στο πρώτο παράδειγμα ένας Machine Learning Engineer σε entry position , full time και με έδρα εταιρίας την Αμερική τοποθετείται σωστά στην medium class. Παρόλο που ο εργαζόμενος είναι σε entry position η ταξινόμηση του στο medium class και οχι στο low μπορεί να δικαιολογηθεί με βάση το γράφημα που μας δείχνει την κατάταξη των μισθών ανα χώρα που βλέπουμε την Αμερική πρώτη με σημαντική διαφορά.

Στο δεύτερο παράδειγμα αλλάζουμε την είσοδο μας σχετικά με την εμπειρία του εργαζομένου όπου του καταχωρούμε EX- executive κρατώντας τις υπόλοιπες τιμές ίδιες. Παρατηρούμε πως το σύστημα τον κατατάσσει στο high class , μια πρόβλεψη που δικαιολογείται με το γράφημα σχετικά με τον μέσο μισθό ανα επίπεδο εμπειρίας όπου οι executives είναι πρώτοι με σημαντική διαφορά. Επίσης βλέπουμε το regression model προβλέπει μισθό περίπου 236.000 που είναι αρκετά πάνω από τον μέσο όρο λόγω των υπόλοιπων τιμών που αποτελούν ευνοϊκές συνθήκες για μεγάλο εισόδημα.

Στο τρίτο παράδειγμα αλλάζουμε την θέση σε MI-Intermediate και την έδρα της εταιρείας σε Ινδία. Το σύστημα πάλι προβλέπει δικαιολογημένη πρόβλεψη και τοποθετεί τον εργαζόμενο στο Low class, καθώς η Ινδία είναι η χώρα με το χαμηλότερο μέσο εισόδημα σε όλες τις θέσεις εργασίας με βάση τα δεδομένα του dataset.

Για την πιο ακριβή προσέγγιση των αποτελεσμάτων έχουμε εισάγει ένα classification report με την βιβλιοθήκη sklearn.metrics όπου μας υπολογίζει μετρικές απόδοσης του μοντέλου μας . Ορισμένες τιμές που που υπολογίζει είναι precision , accuracy , support , f1-score κτλ

```
Classification Report:
              precision    recall  f1-score   support

        high       0.63      0.24      0.35       164
         low       0.85      0.64      0.73       164
      medium       0.65      0.90      0.75       371

    accuracy                           0.68       699
   macro avg       0.71      0.59      0.61       699
weighted avg       0.69      0.68      0.65       699
```

Το μοντέλο μας έχει ακρίβεια προβλέψεων 68% που μεταφράζεται σε 68 σωστές προβλέψεις ανα 100. Συνολικά, το μοντέλο φαίνεται να έχει καλή απόδοση για τις κατηγορίες "low" και "medium", αλλά δυσκολεύεται να προβλέψει σωστά την κατηγορία "high".

7. ***ΣΥΜΠΕΡΑΣΜΑΤΑ***

Με βάση τις εξόδους που λάβαμε τα αποτελέσματα μπορούν να χαρακτηριστούν επιτυχή, βέβαια με την βοήθεια των εργαλείων υπολογισμού μετρικών μπορούμε να διαπιστώσουμε πως τα συστήματα μπορούν να βελτιωθούν σημαντικά.

Σε γενικές γραμμές δεδομένου του σκοπού μας για πρόβλεψη και ταξινόμησή των μισθών σε διεθνή κλίμακα τα αποτελέσματα των μετρικών εμφανίζονται της τάξης του 68% λόγω της πολυπλοκότητας του. Για παράδειγμα το dataset περιέχει πολλούς τίτλους εργασίας σε σχέση με το συνολικό μέγεθος της όποτε λόγω περιορισμένων εγγραφών σε κάποιες θέσεις η χώρες η ακρίβεια πέφτει. Μια πιο περιορισμένη ανάλυση των μοντέλων μονο σε χώρες, όπως η Αμερική , θα ήταν αρκετά πιο επιτυχημένη λόγω του πλήθους των δεδομένων που κατέχουμε.

Για να βελτιωθούν τα συστήματα και να ανέβει η ακρίβεια των αποτελεσμάτων μπορούν να γίνουν αρκετές αλλαγές. Μια απο αυτές είναι η πρόσθεση επιπλέον δεδομένων στο dataset μας καθώς αυτό θα βελτιώσει το training model μας και θα ξεκαθαρίσει δεδομένα όπου υπάρχουν λίγες εγγραφές και μπορεί τα αποτελέσματα να μην αναγνωρίζονται σωστά.

Ακόμα μπορούν να διερευνηθούν περισσότεροι αλγόριθμοι για τις τεχνικές που ακολουθήσαμε για εύρεση του καταλληλότερου για το dataset μας , όπως support vector machines- SVM ή Decision Trees για το classification.

Τέλος θα μπορούσε να χρησιμοποιηθεί ένα εκτενές correlation matrix που θα μας εμφάνιζε σε καλύτερο βαθμό την βαρύτητα των στηλών που επηρεάζουν την στήλη του μισθού. Με αυτό το τρόπο η επιλογή των στηλων για το training model θα γινόταν με βαση την βαρύτητα και μπορεί να ανέβαζε την ακρίβεια των συστημάτων.

Βιβλιογραφία

1. *Data Science Salaries 2023* (n.d.). Ιούλιος 1, 2023, από https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023