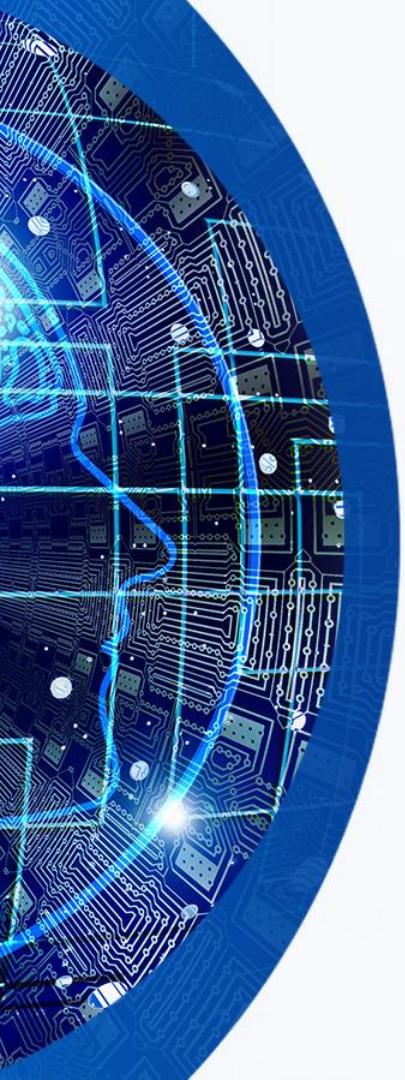
Data Challenge

Team

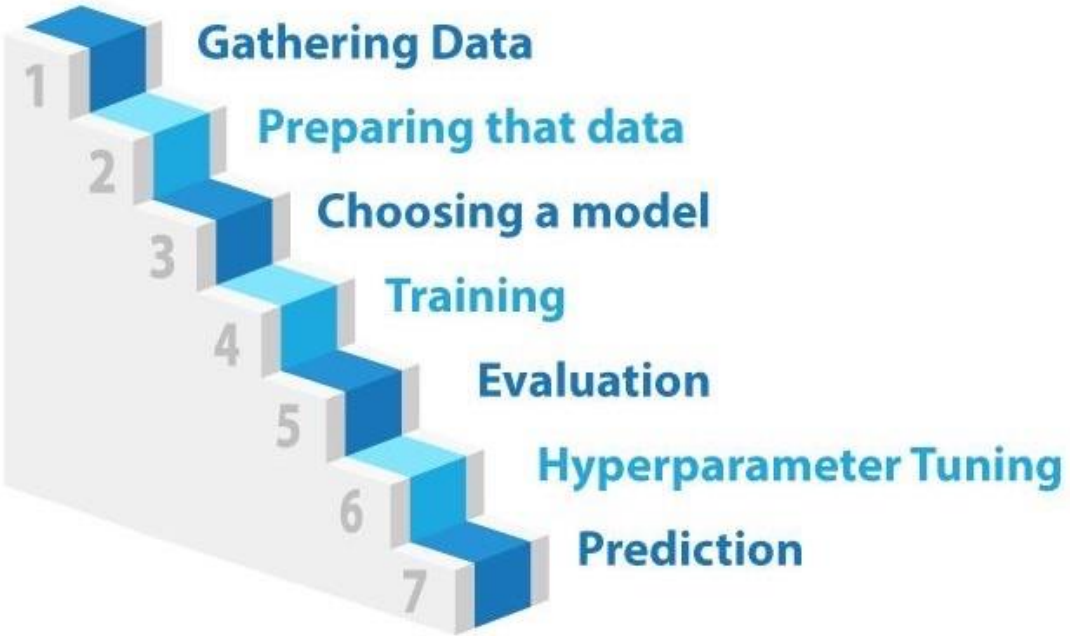
Gkountoumas Filippos
Legkas Sotiris
Papadopoulos Nikolaos

Domain Name Classification Challenge

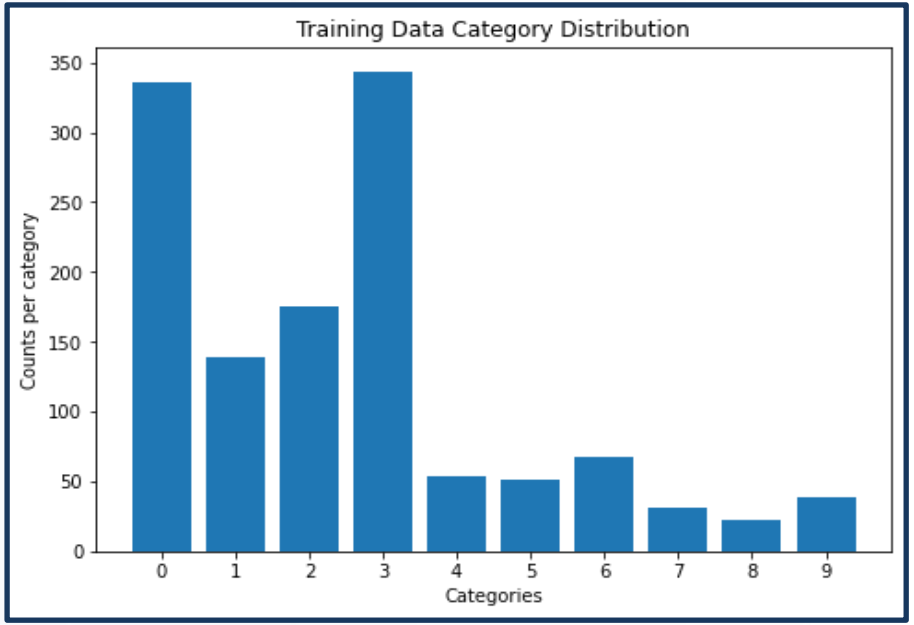
- Study and Apply ML/ Data mining Techniques
- Assign Domain Names into a Predefined Category
- Textual data & Web Graph
- Goal : Predict the class of unlabeled domains



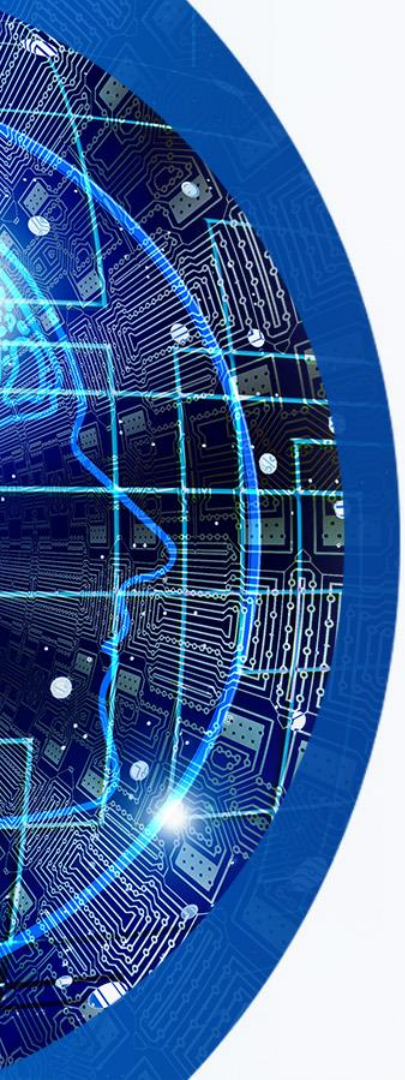
Project Pipeline



Data Exploration



- Unbalanced Data
- Major categories “0”&”3”
- Categories “7”,”8”,”9” less than 25 obs

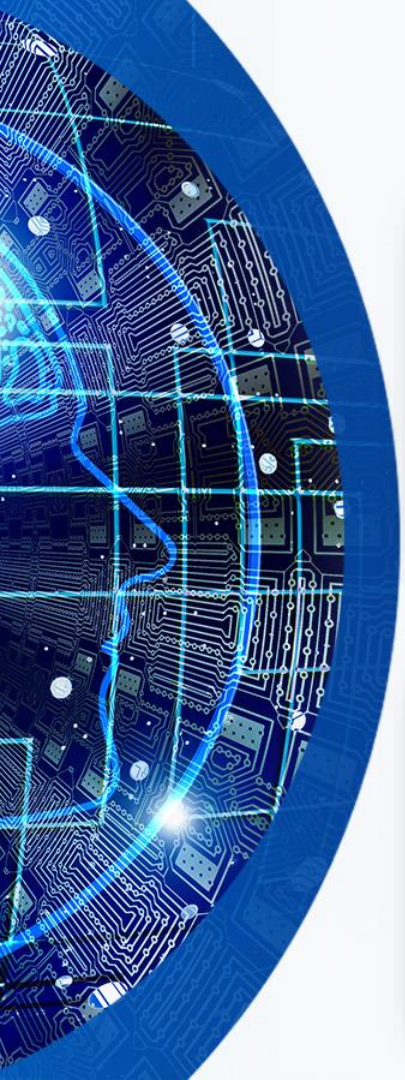


Data Preprocessing (Data Cleaning)

- **Removal of**
  HTML tags
  Punctuation and symbols
  Numbers
  Greek stop words
  Words under 4 letters
  Multiple spaces

- **Further preprocessing**
  Lowercasing the text
  Accent stripping
  Text length shortening (2500 words per text)
  Stemming (using Greek Stemmer)

Data Preprocessing (Data Cleaning)

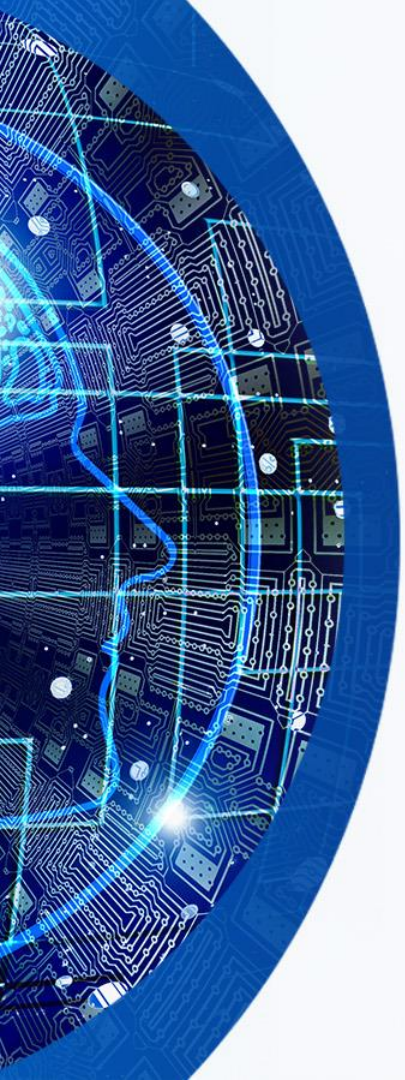


'http://kollintzas.gr/%CE%94%CE%99%CE%91%CE%93%CE%A9%CE%9D%CE%99%CE%A3%CE%9C%CE%9F%CE%A3%20%CE%95%CE%9A%CE%A0%CE%91%CE%99%CE%94%CE%95%CE%A5%CE%A4%CE%99%CE%9A%CE%A9%CE%9D.htm *#* ευρώ\n Καταβολή διδάκτρων Στη η διδακτική συνάντηση\nΓια δηλώσεις συμμετοχής στα δωρεάν δοκιμαστικά μαθήματα απαιτείται η δήλωση συμμετοχής τουλάχιστον τρείς ημέρες πριν την ημερομηνία έναρξης\nΣτους υποψηφίους θα δοθούν εκτός των επιμελημένων δακτυλογραφημένων σημειώσεων που συνοδεύουν την κάθε διδακτική συνάντηση συνολικά ερωτήσεις πολλαπλών επιλογών \nΑΝΑΚΟΙΝΩΣΕΙΣ ΣΧΕΤΙΚΑ ΜΕ ΤΗΝ ΠΟΡΕΙΑ ΤΟΥ ΠΡΟΣΕΧΟΥΣ ΔΙΑΓΩΝΙΣΜΟΥ \nΔιαγωνισμός ΑΣΕΠ Εκπαιδευτικών Νέες Δηλώσεις κ Υπουργού \nΕφόσον οι πιστωτές της χώρας συναινέσουν στο σχέδιο του υπουργού Παιδείας τότε θα ανοίξει ο δρόμος για τη διενέργεια διαγωνισμού ΑΣΕΠ για την πρόσληψη εκπαιδευτικών στα σχολεία Όπως αποκαλύπτει ο κ Λοβέρδος στην Η στόχος είναι ο διαγωνισμός να διενεργηθεί τον χειμώνα και συγκεκριμένα ένα τον Δεκέμβριο του ή τον Ιανουάριο του και να αφορά πέντε εγάλες κατηγορίες εκπαιδευτικών φιλολόγους αθηματικούς φυσικούς δασκάλους και νηπιαγωγούς \nΣύμφωνα με υπολογισμούς του υπουργείου Παιδείας για την ομαλή έναρξη των σχολείων το απαιτούνται περίπου τις προσλήψεις τακτικού προσωπικού Ση ειώνεται ότι ο τελευταίος διαγωνισμός του ΑΣΕΠ είχε διενεργηθεί το καλοκαίρι από αυτόν δεν έχουν ακόμη διοριστεί επιτυχόντες εκπαιδευτικοί \nΟ κ Λοβέρδος εκφράζει αισιοδοξία για το αποτέλεσμα των διαπραγματεύσεων που θα έχει με την τρόικα σε τεχνικό επίπεδο βασιζόμενος ότι η κατάσταση στον χώρο της εκπαίδευσης έχει ξεπεράσει τα όρια Το σύστημα κάλυψης των κενών στα σχολεία εναπληρωτές πεθαίνει Από τη μία στερεύουν οι πηγές χρηματοδότησης των αναπληρώσεων έσω του Προγράμματος Δημοσίων Επενδύσεων και του ΕΣΠΑ κι από την άλλη εξαντλούνται τα έτρα πολιτικής επισημαίνει ο κ Λοβέρδος αποκλείοντας το ενδεχόμενο να προχωρήσει σε νέο κύκλο συγχωνεύσεων καταργήσεων σχολικών ονάδων \nΆλλωστε όπως αναφέρει ο ίδιος από το σχέδιο Αθηνά για την αναδιάρθρωση των πανεπιστημίων και ΤΕΙ της χώρας δεν εξοικονομήθηκε ούτε ένα ευρώ παρότι η κυβέρνηση στόχευε να κάνει οικονομία της τάξεως των εκατ ευρώ Αυτό αποκάλυψε ο αναπληρωτής υπουργός Οικονομικών Χρήστος Σταϊκούρας σε πρόσφατη σύσκεψη που έγινε στο Μέγαρο Μαξίμου για τα οικονομικά των ΑΕΙ παρουσία του πρωθυπουργού και του αντιπροέδρου της κυβέρνησης \nΤο αίτημα στην τρόικα\nΣε μια προσπάθεια να πείσει τη

'ευρω καταβολ διδακτρ διδακτικ συναντης δηλως συμμετοχ δωρε δοκιμαστικ μαθηματ απαιτειται δηλως συμμετοχ τουλαχιστ ημερ πριν η μερομηνι εναρξ υποψηφι δοθ εκτος επιμελ δακτυλογραφ σημειως συνοδευ καθ διδακτικ συναντης συνολικ ερωτης πολλαπλ επιλογ ανακοινως σχετικ πορει προσεχ διαγωνισμ διαγωνισμ ασεπ εκπαιδευτικ δηλως υπουργ εφος πιστωτ χωρ συναινες σχεδι υπουργ παιδει τοτε ανοιξ διενεργει διαγων ασεπ προσληψ εκπαιδευτικ σχολει οπως αποκαλυπτ λοβερδ στην στοχ εινα διαγων διενεργηθ συγκεκρ δεκ βρι ι ανουαρι αφορ πεντε εγαλ κατηγορι εκπαιδευτικ φιλολογ ατικ φυσικ δασκαλ νηπιαγωγ φων υπολογ υπουργει παιδει εναρξ σχολ απαιτουν τα περι προσληψ τακτικ προσωπικ ειωνετα τελευται διαγων ασεπ ειχ διενεργηθ αυτ διοριστ επιτυχοντ εκπαιδευτικ λοβερδ εκφραζ αισιοδοξι αποτελ διαπραγ ατευσ τροικα τεχνικ επιπεδ βασιζ οπως γεγονο καταστας χωρ εκπαιδευσ ξεπερας ορι συστ καλυψ κεν σχολει αναπληρωτ πεθαιν στερευ πηγ ατοδοτης αναπληρωσ προγρ οσι επενδυσ εσπα αλλ εξαντλουντα ετρ πολιτικ επισ αιν λοβερδ αποκλεισ ενδεχ προχωρησ κυκλ συγχωνευσ καταργησ σχολικ ονοδ αλλωστε οπως αναφερ ιδι σχεδι αθην αναδιαρθρωσ πανεπιστ χωρ εξοικονομηθηκ ουτ ευρω παρομ κυβερνησ στοχευ καν οικονομ ταξ εκατ ευρω αυτ αποκαλυψ αναπληρωτ υπουργ οικονομ χρηστ σταικουρ προσφατ συσκεψ εγιν μεγαρ μαξι οικονομ παρουσι πρωθυπουργ αντιπροεδρ κυβερνησ αιτ στην τροικα προσπαθει πεισ τροικα αναγκ προσληψ προσωπικ στην εκπαιδευσ ανδρε λοβερδ δως εντολ οδι υπηρεσι υπουργει περιφερειακ διευθυνσ εκπαιδευσ καθως ινστιτουτ εκπαιδευτικ πολιτικ προχωρησ πληρ χαρτογραφησ εκπαιδευτικ αθητικ δυν σχολικ ονοδ πρωτοβαθμ δευτεροβαθμ εκπαιδευσ χωρ μαθητ ταξ πλαισι αυτ συγκεντρωνοντα αυτ διαστ αναλυτικ στοιχει αφορ εταξ αλλ αριθμ αθητ ταξ αναλογι τους διδασκοντ ιδι βαρυτητ δινετα στην εικονομ παρουσιαζ σχολει νησιωτικ χωρ πρεπ γιν κατανοητ στην τροικα ιδιαιτεροτητ ελλαδ χωρ νησιωτικ ελλειψ εκπαιδευτικ ιουργ εγαλ προβλ στην ευρυθμ λειτουργι σχολικ ονοδ ικρ νησι χωρ υπουργ στην υπουργ παιδει τονιζ αναγκ σχεδι στρατηγικ στην εκπαιδευσ βασιζετα νησιωτικ ιδιαιτεροτητ χωρ λοβερδ υστερα συνεννοησ πρωθυπουργ αντων αντιπροεδρ κυβερνησ ευαγγελ βενιζελ σκοπευ υποβαλ σεπτ βρι στην τροικα αιτ εξαιρεθ εκπαιδευσ κανον προσληψ αποφασ ερωσ σχετικ υπουργ διοικητικ μεταρρυθμ κυριακ μητσοτακ οποι φανιζετα θετικ αφορ υπουργει διοικητικ μεταρρυθμ αφορ υπουργει παιδει συζητησ ευρωπαικ επιτροπ ειων υπουργ προγραμμ θεματικ ενοτητ μαθημ κωδικ δικασ

Original Text

Processed Text



Feature Engineering

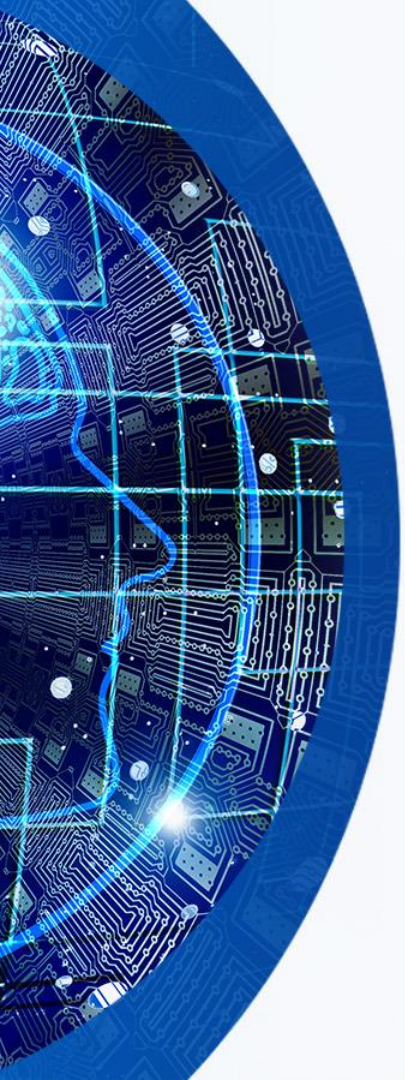
- **TF-IDF**
  Only kept words occurring more than 50 and less than 300 times
  After the data transformation only used the 20.000 most frequent words

- **Greek Word Embeddings**
  Used fastext embeddings from AUEB

- **Dimensionality Reduction**
  Used Truncated SVD to reduce feature dimensionality to 128

- **Graph creation**
  Created a Graph with all node connections from the given file

- **Node2Vec**
  Implemented by combining StellarGraph's random walk generator with the word2vec algorithm from Gensim



Model Creation

- **GCN**

  Approach for semi-supervised learning on graph-structured data.
  Spatial Convolution works on a local neighbourhood of nodes and understands

  Model Structure

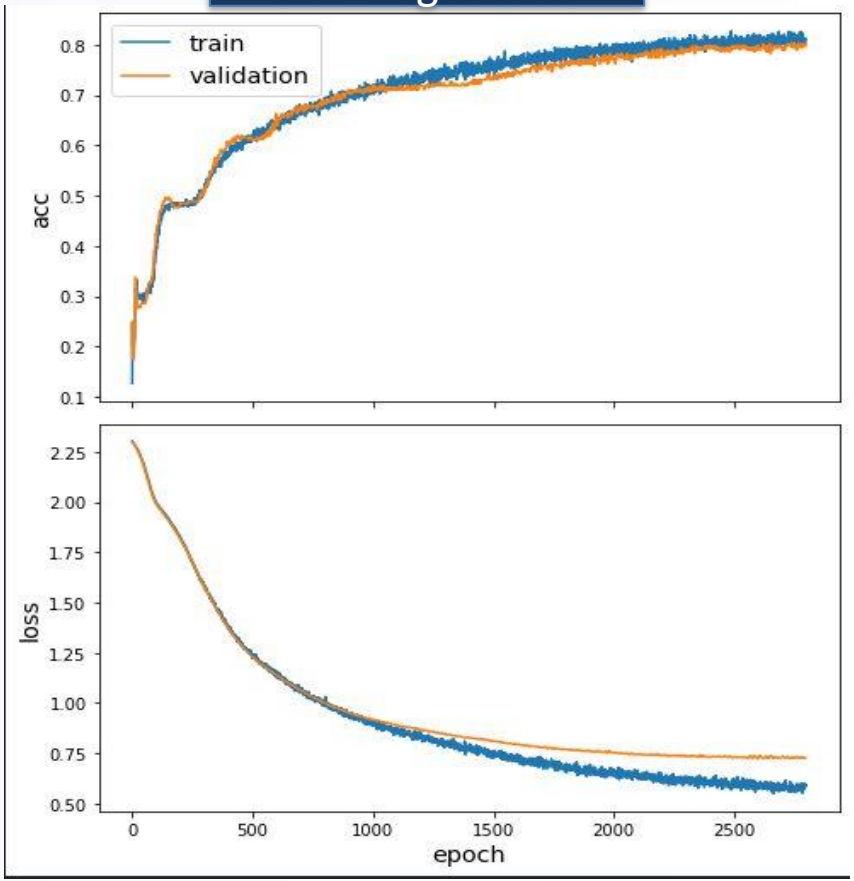
  - 2 Convolution Graph Layers (32 units each, activation : ReLU)
  - 2 Dropout Layers (Dropout Rate 0.5 to avoid overfitting)
  - 1 Dense Layer (10 units, activation : SoftMax)

  Model fit

  - Loss : Categorical Cross Entropy
  - Optimizer : Adam
  - Learning Rate : 10e-3
  - Split : Stratified to maintain class balance

Model Evaluation

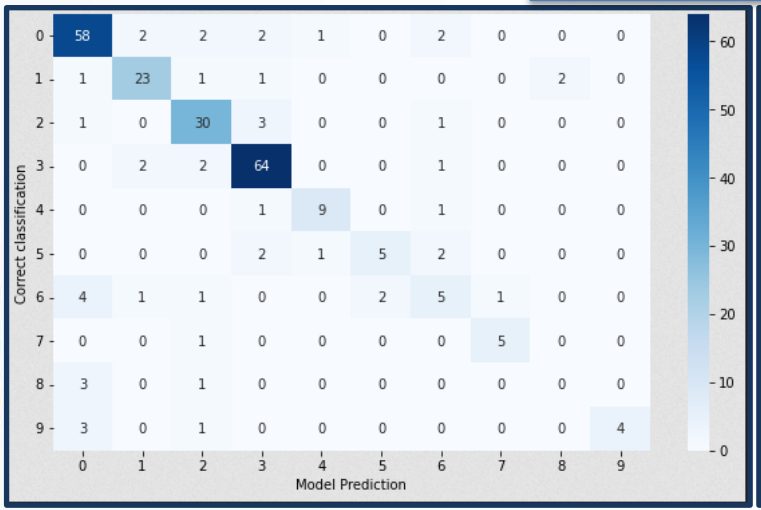
Learning Curves

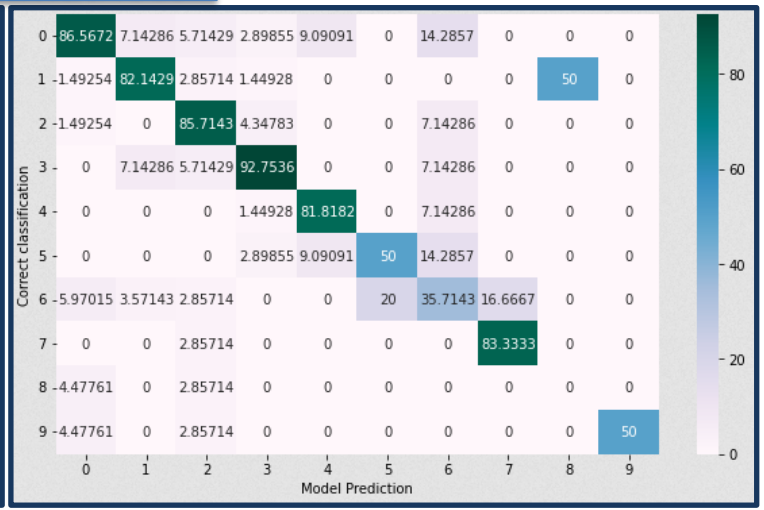


- No overfitting occurred
- Excessive fluctuation might be a sign of unrepresentative dataset

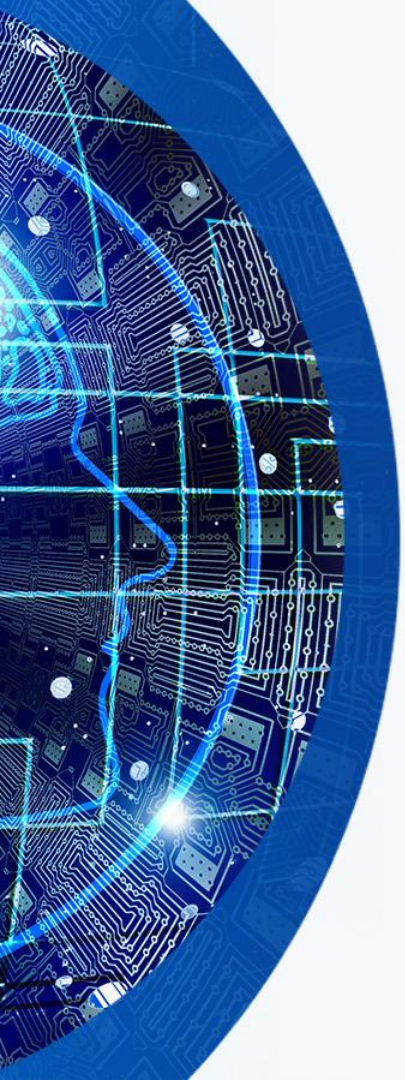
Error Analysis

Confusion Matrices





- Domains belonging to the most common categories, get classified with higher accuracy
- Categories with less observation are way more difficult to classify Especially categories that even human annotators would have difficulties classifying



Thank you