ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Διαχείριση δεδομένων και ανάλυση με μεθόδους μηχανικής μάθησης με χρήση του περιβάλλοντος Microsoft Azure

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Μαμουλέλλη Απόστολου

Επιβλέπων:

Δρ. Λουκής Ευριπίδης, Καθηγητής

Μέλη εξεταστικής επιτροπής:

Δρ. Διαμαντοπούλου Βασιλική, Επίκουρη Καθηγήτρια

Δρ. Καρύδα Μαρία, Καθηγήτρια

Σάμος, *Μάρτιος* 2024

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Η παρούσα Διπλωματική Εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2023-2024 στα πλαίσια του Προπτυχιακού Προγράμματος Σπουδών του τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων που αποτελεί μέρος της Πολυτεχνικής Σχολής του Πανεπιστημίου Αιγαίου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή της διπλωματικής μου εργασίας Δρα Ευριπίδη Λουκή, για τη συνεχή υποστήριξη, τις συμβουλές και την καθοδήγησή του κατά τη διάρκεια εκπόνησης αυτής της εργασίας.

Ακόμη, θέλω να ευχαριστήσω ιδιαιτέρως την οικογένειά μου για τη στήριξή τους και την κατανόησή τους καθ' όλη τη διάρκεια των σπουδών μου.

Τέλος, θα ήθελα να ευχαριστήσω από καρδιάς τους φίλους μου, που αν και απομακρυσμένοι φυσικά, η ενθάρρυνση, οι συμβουλές και η εμπιστοσύνη που μου έδειξαν με έκαναν να αισθανθώ πιο δυνατός και πιο αποφασισμένος να ολοκληρώσω αυτό το σημαντικό εγχείρημα.

Περίληψη

Η παρούσα εργασία εστιάζει στην ανάλυση και την εφαρμογή προηγμένων τεχνολογιών διαχείρισης δεδομένων και μηχανικής μάθησης στο περιβάλλον του Microsoft Azure. Εξερευνούμε τη χρήση διαφόρων υπηρεσιών, συμπεριλαμβανομένων των Data Warehouse, Data Lake, Data Lakehouse, Serverless SQL Pool και Dedicated SQL Pool, για την αποθήκευση, την ανάλυση και την ανάκτηση δεδομένων. Στη συνέχεια, παρουσιάζουμε τρία διαφορετικά σενάρια εφαρμογής, τα οποία καλύπτουν την αναζήτηση και ανάλυση δεδομένων σε Data Lakes με χρήση Serverless SQL Pool, καθώς και την εκπαίδευση προγνωστικών μοντέλων παλινδρόμησης και ταξινόμησης με τη χρήση του Azure Machine Learning. Μέσα από αυτά τα σενάρια, εξερευνούμε την ικανότητα του Microsoft Azure να παρέχει ένα ολοκληρωμένο περιβάλλον για τη διαχείριση και την ανάλυση δεδομένων, καθώς και την ανάπτυξη προηγμένων μοντέλων μηχανικής μάθησης για προβλέψεις και ταξινομήσεις. Αυτή η εργασία προσφέρει μια διαφορετική οπτική στη σύγχρονη διαχείριση δεδομένων και την μηχανική μάθηση μέσω των υπηρεσιών του Microsoft Azure.

Abstract

This paper focuses on the analysis and implementation of advanced data management and machine learning technologies in the Microsoft Azure environment. We explore the use of various services, including Data Warehouse, Data Lake, Data Lakehouse, Serverless SQL Pool, and Dedicated SQL Pool, for data storage, analysis, and retrieval. We then present three different application scenarios, covering data search and analysis in Data Lakes using Serverless SQL Pool, as well as training predictive regression and classification models using Azure Machine Learning. Through these scenarios, we explore the ability of Microsoft Azure to provide an integrated environment for data management and analysis, as well as the development of advanced machine learning models for prediction and classification. This paper offers a different perspective on modern data management and machine learning through Microsoft Azure services.

Περιεχόμενα

1. Εισαγωγή

Η διαχείριση των δεδομένων αποτελεί κρίσιμη πτυχή για την επιτυχή λειτουργία και την καινοτομία σε οργανισμούς και επιχειρήσεις κάθε μεγέθους. Με την ανάπτυξη και την εξέλιξη των τεχνολογιών δεδομένων, όπως οι διαδικτυακές υπηρεσίες και οι αισθητήρες, η ποσότητα και η ποικιλία των δεδομένων που διαχειρίζονται οι οργανισμοί έχει εκθειάσει την ανάγκη για αποτελεσματικές λύσεις αποθήκευσης, ανάκτησης και ανάλυσης δεδομένων.

Στο πλαίσιο αυτό, οι υπηρεσίες του Microsoft Azure έχουν αναδειχθεί ως ένας κορυφαίος πάροχος υπηρεσιών cloud computing που προσφέρει πληθώρα λύσεων για τη διαχείριση δεδομένων και την ανάπτυξη εφαρμογών. Στο πλαίσιο αυτής της εργασίας, εστιάζουμε στην ανάλυση και την εφαρμογή των προηγμένων λύσεων διαχείρισης δεδομένων και μηχανικής μάθησης που προσφέρει το Microsoft Azure.

Η εργασία πραγματεύεται τη χρήση διαφόρων υπηρεσιών του Microsoft Azure, όπως το Data Warehouse, το Data Lake, το Data Lakehouse, το Serverless SQL Pool και το Dedicated SQL Pool, για την αποθήκευση, τη διαχείριση και την ανάκτηση δεδομένων. Αντικείμενο της παρούσας εργασίας είναι αρχικά η γνωριμία με την επιστήμη των δεδομένων, την επιχειρησιακή αναλυτική, την επιχειρησιακή ευφυΐα και την προγνωστική αναλυτική, την κατανόηση των όρων αυτών και της κατανόησης των βασικών γνωρισμάτων κάθε μίας από αυτές. Στην συνέχεια, μέσω της υλοποίησης μίας σειράς σχετικών σεναρίων, στόχο μας αποτελεί να κατανοήσουμε την χρησιμότητα των προγραμμάτων αυτών στην καθημερινότητα μιας επιχείρησης και στην βοήθεια που μπορεί να προσφερθεί για την σωστή λήψη των αποφάσεων.

Συγκεκριμένα στο 2ο κεφάλαιο γίνεται μια αναλυτική παρουσίαση όλων των χρήσιμων και σημαντικών όρων, όπως η επιστήμη των δεδομένων, η επιχειρησιακή αναλυτική, οι υπολογιστικές υπηρεσίες, η τεχνητή νοημοσύνη και η μηχανική μάθηση, ο ορισμός αυτών και οι βασικές τους έννοιες και κατηγορίες. Στο 3ο κεφάλαιο γίνεται αναφορά στην μεθοδολογία που ακολουθήθηκε για την υλοποίηση των σεναρίων. Στο 4ο κεφάλαιο παρουσιάζονται τρία σενάρια, κάθε ένα από τα οποία έχει μεγαλύτερο βαθμό δυσκολίας στην υλοποίηση του από το προηγούμενο. Για την δημιουργία αυτών των σεναρίων χρησιμοποιείται το εργαλείο Azure της Microsoft.

Συνοπτικά, το πρώτο σενάριο μελετά την εκτέλεση μιας διερευνητικής ανάλυσης δεδομένων χρησιμοποιώντας Serverless SQL Pool. Ο σκοπός μας είναι να συνδυάσουμε διαφορετικά σύνολα δεδομένων για την άμεση και εύκολη εξαγωγή πληροφοριών ή συμπερασμάτων. Στο δεύτερο σενάριο θα δούμε πως μπορούμε να εκπαιδεύσουμε ένα μοντέλο παλινδρόμησης (Regression Model) για μια εταιρεία πωλήσεων αυτοκινήτων που θέλει να χρησιμοποιήσει τα χαρακτηριστικά ενός αυτοκινήτου για να προβλέψει την πιθανή τιμή πώλησης. Στο τρίτο και τελευταίο σενάριο θα δούμε πώς μπορούμε να εκπαιδεύσουμε ένα μοντέλο ταξινόμησης (Classification Model) για την πρόβλεψη της πιθανής εμφάνισης διαβήτη σε ασθενής, σύμφωνα με το ιατρικό τους ιστορικό.

Στο 5ο κεφάλαιο κάνουμε μια σύνοψη όσων είδαμε στο προηγούμενο κεφάλαιο, αναλύουμε τα αποτελέσματα των σεναρίων που υλοποιήσαμε και προσδιορίζουμε τις δυνατότητες και τα πλεονεκτήματα της χρήσης του Microsoft Azure στον τομέα της διαχείρισης δεδομένων

και της μηχανικής μάθησης. Τέλος στο 6ο κεφάλαιο παρουσιάζουμε τα συμπεράσματα μας, που αφορούν τόσο το περιβάλλον του Azure όσο και την χρησιμότητα του χώρου στην ανάπτυξη των επιχειρήσεων.

Συνοψίζοντας, η συγκεκριμένη εργασία παρέχει μια ολοκληρωμένη επισκόπηση των προηγμένων τεχνολογιών διαχείρισης δεδομένων και μηχανικής μάθησης στο Microsoft Azure, καθώς και την εφαρμογή τους μέσω συγκεκριμένων σεναρίων.

2. Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή στην Επιστήμη των Δεδομένων

Η επιστήμη των δεδομένων είναι ένας διεπιστημονικός τομέας που ασχολείται με την εξαγωγή γνώσης και διορατικότητας από δομημένα και αδόμητα δεδομένα. Η πρακτική της απόκτησης χρήσιμων πληροφοριών από τα δεδομένα είναι ιδιαίτερα σημαντική για μεγάλα ή πολλά δεδομένα (Big Data). Καθημερινά δημιουργούνται ή συλλέγονται petabytes δομημένων και μη δομημένων δεδομένων από πολλές πηγές εντός ή και εκτός των οργανισμών, των επιχειρήσεων και των ιδιωτών. Αυτό έχει σαν αποτέλεσμα, ο κόσμος να είναι πλούσιος σε δεδομένα αλλά φτωχός σε πληροφορίες που προκύπτουν από αυτά. Πιο συγκεκριμένα, το petabyte είναι μια μονάδα αποθήκευσης ψηφιακών πληροφοριών, όπου σε μια αίσθηση της κλίμακας το 1 petabyte ισοδυναμεί με χίλια (≈1.024) TB ή αντίστοιχα με ένα εκατομμύριο (≈1.024.000) GB. Επομένως, τα petabyte χρησιμοποιούνται συνήθως για να περιγράψουν πολύ μεγάλες ποσότητες δεδομένων, όπως αυτές στην ανάλυση μεγάλων δεδομένων, στην επιστημονική έρευνα σε μεγάλη κλίμακα ή σε συστήματα αποθήκευσης που βασίζονται στο σύννεφο. Η επιστήμη των δεδομένων (Data Science) συνδυάζει διάφορες τεχνικές και μεθόδους από τη στατιστική, τα μαθηματικά, την επιχειρησιακή έρευνα, την επιστήμη των υπολογιστών και την ανάλυση δεδομένων με σκοπό να αναλύσει και να ερμηνεύσει πολύπλοκα δεδομένα. Ουσιαστικά, ο πρωταρχικός στόχος της επιστήμης δεδομένων είναι να καταστήσει τα δεδομένα χρήσιμα, κατατοπιστικά και εφαρμόσιμα. Αν και ο όρος επιστήμη δεδομένων είναι νέος στις επιχειρήσεις, υπάρχει από το 1960, όταν χρησιμοποιήθηκε για πρώτη φορά από τον Peter Naur για να αναφερθεί σε μεθόδους επεξεργασίας δεδομένων στην επιστήμη των υπολογιστών. Από τα τέλη της δεκαετίας του 1990 αξιοσημείωτοι στατιστικολόγοι, όπως ο C.F. Jeff Wu και ο William S. Cleveland, χρησιμοποιούν επίσης τον όρο επιστήμη των δεδομένων, έναν κλάδο που θεωρούν ως το ίδιο ή ως επέκταση της στατιστικής (Barga, Roger, et al, 2015).

Συνοψίζοντας, η επιστήμη των δεδομένων αναδεικνύεται ως κρίσιμος παράγοντας στο σύγχρονο επιχειρηματικό περιβάλλον. Με την ικανότητά της να αντλεί πληροφορίες από ποικίλες πηγές και να αναλύει τα δεδομένα με εξειδικευμένες μεθόδους, δίνει στις επιχειρήσεις τη δυνατότητα όχι μόνο να κατανοούν τις τρέχουσες συνθήκες, αλλά και να προβλέπουν τις μελλοντικές εξελίξεις. Από την περιγραφική ανάλυση μέχρι την προγνωστική διάσταση, η επιστήμη των δεδομένων ανοίγει νέους ορίζοντες για τη λήψη αποφάσεων με βάση την ενδελεχή αξιοποίηση των διαθέσιμων πληροφοριών. Σε έναν κόσμο όπου η ποσότητα των δεδομένων συνεχώς αυξάνεται, η ικανότητα εκμετάλλευσης της επιστήμης των δεδομένων αποτελεί κρίσιμο εργαλείο για τη διατήρηση της ανταγωνιστικότητας και την καλλιέργεια της καινοτομίας σε κάθε τομέα. Καθώς προβαίνουμε σε μια ολοκληρωμένη επισκόπηση της επιστήμης των δεδομένων, η μετάβαση προς την επιχειρηματική αναλυτική γίνεται αναπόφευκτη. Εκείνο που καθιστά την επιχειρηματική αναλυτική ακόμα πιο σημαντική είναι η ικανότητά της να ενσωματώνει τα αποτελέσματα της επιστήμης των δεδομένων σε στρατηγικές και διαδικασίες. Ενώ η επιστήμη των δεδομένων παρέχει την αναγκαία κατανόηση και πρόβλεψη, η επιχειρηματική αναλυτική προσφέρει τη δομημένη διαδικασία για την αξιοποίηση αυτών των γνώσεων προς όφελος της επιχείρησης. Έτσι, η

συνεργασία μεταξύ αυτών των δύο πεδίων ενισχύει την ικανότητα λήψης αποφάσεων και την ανταπόκριση στις απαιτήσεις του σύγχρονου επιχειρηματικού περιβάλλοντος.

2.2 Επιχειρησιακή Αναλυτική

Η επιχειρηματική ανάλυση είναι η διαδικασία χρήσης τεχνικών ανάλυσης δεδομένων και στατιστικών τεχνικών για τη διερεύνηση, ανάλυση και ερμηνεία ιστορικών και τρεχόντων δεδομένων με στόχο την απόκτηση γνώσεων και την υποστήριξη της λήψης αποφάσεων βάσει δεδομένων σε μια επιχείρηση ή έναν οργανισμό. Ο στόχος της επιχειρηματικής ανάλυσης είναι να βοηθήσει τις επιχειρήσεις να λαμβάνουν τεκμηριωμένες και βασισμένες σε γεγονότα αποφάσεις για τη βελτίωση των λειτουργιών, την αύξηση της αποδοτικότητας, τον εντοπισμό ευκαιριών και την αντιμετώπιση των προκλήσεων. Τα αναλυτικά στοιχεία που προκύπτουν μπορεί να είναι χρήσιμα για ανθρώπινες αποφάσεις ή μπορεί να οδηγούν πλήρως σε αυτοματοποιημένες αποφάσεις (Davenport, Thomas and Harris, 2007). Η επιχειρησιακή αναλυτική χωρίζεται συνήθως σε 3 επίπεδα, παρόλο που διάφοροι συγγραφείς πολλές φορές στις αναλύσεις τους την χωρίζουν και σε άλλα μικρότερα επίπεδα. Στον παρόν διαχωρισμό θα δούμε τα 3 επίπεδα που είναι αποδεκτά από όλες τις αναλύσεις που γίνονται (Davenport, Thomas and Harris, 2007).

- **Περιγραφική Αναλυτική (Descriptive Analysis)**: Η περιγραφική αναλυτική είναι μια στατιστική μέθοδος που χρησιμοποιείται για την αναζήτηση και τη σύνοψη ιστορικών δεδομένων προκειμένου να προσδιοριστούν μοτίβα και επικεντρώνεται μόνο σε ό, τι έχει ήδη συμβεί σε μια επιχείρηση. Η περιγραφική ανάλυση χρησιμοποιείται για να εξηγήσει τι συμβαίνει σε μια δεδομένη κατάσταση και χρησιμοποιεί δύο βασικές μεθόδους για να ανακαλύψει ιστορικά δεδομένα, τη συλλογή δεδομένων και την εξόρυξη δεδομένων ώστε. Η συγκέντρωση δεδομένων είναι η διαδικασία συλλογής και οργάνωσης δεδομένων για τη δημιουργία συνόλων δεδομένων που είναι διαχειρίσιμα. Στη συνέχεια, αυτά τα σύνολα χρησιμοποιούνται στη φάση εξόρυξης δεδομένων όπου τα πρότυπα, οι τάσεις και το νόημα αναγνωρίζονται και έπειτα παρουσιάζονται με κατανοητό τρόπο χρησιμοποιώντας κατάλληλες μεθόδους απεικόνισης (Visualization). Η παρουσίαση τους γίνεται με εξειδικευμένες αναφορές που περιλαμβάνουν ιστογράμματα, bar charts, line charts, pies, άλλου είδους διαγράμματα ή πίνακες.
- **Προγνωστική Αναλυτική (Predictive Analytics)**: Η προγνωστική αναλυτική είναι μια στατιστική μέθοδος που χρησιμοποιεί αλγόριθμους στατιστικής και μηχανική μάθηση για τον εντοπισμό τάσεων στα δεδομένα και την πρόβλεψη μελλοντικών συμπεριφορών. Πρακτικά χρησιμοποιείται για την πρόβλεψη της πιθανότητας ενός αβέβαιου αποτελέσματος έτσι ώστε να προετοιμάσει πολλές πτυχές μιας επιχείρησης ή ενός οργανισμού, συμπεριλαμβανομένου του καθορισμού ρεαλιστικών στόχων, του αποτελεσματικού σχεδιασμού, της διαχείρισης των προσδοκιών απόδοσης και της αποφυγής κινδύνων. Η στατιστική και η μηχανική μάθηση προσφέρουν εξαιρετικές τεχνικές για την πρόβλεψη, όπως τα νευρωνικά δίκτυα, τα δέντρα αποφάσεων, τα τυχαία δάση, τα ενισχυμένα δέντρα αποφάσεων, την προσομοίωση Μόντε Κάρλο και την παλινδρόμηση. Για παράδειγμα, οι

αλγόριθμοι μηχανικής μάθησης λαμβάνουν τα υπάρχοντα δεδομένα και προσπαθούν να συμπληρώσουν τα δεδομένα που λείπουν με τις καλύτερες δυνατές υποθέσεις για να κάνουν τις προβλέψεις.

- **Καθοδηγητική Αναλυτική (Prescriptive Analysis)**: Η καθοδηγητική αναλυτική είναι μια στατιστική μέθοδος που χρησιμοποιείται για τη δημιουργία συστάσεων και τη λήψη αποφάσεων με βάση τα υπολογιστικά ευρήματα των αλγοριθμικών μοντέλων. Συνήθως, συνδυάζει ένα προγνωστικό μοντέλο με επιχειρηματικούς κανόνες, όπως η απόρριψη μιας συναλλαγής εάν η πιθανότητα απάτης είναι πάνω από ένα συγκεκριμένο όριο. Για παράδειγμα, μπορεί να προτείνει το καλύτερο τηλεφωνικό πρόγραμμα που πρέπει να προσφέρεται σε έναν συγκεκριμένο πελάτη, ή με βάση τη βελτιστοποίηση, μπορεί να προτείνει την καλύτερη διαδρομή για τα φορτηγά διανομής της εταιρίας. Οι τεχνικές που μπορούν να χρησιμοποιηθούν για την πραγματοποίηση καθοδηγητικής ανάλυσης περιλαμβάνουν τα δέντρα αποφάσεων, ο γραμμικός και ο μη γραμμικός προγραμματισμός, η προσομοίωση Monte Carlo ή η θεωρία παιγνίων από τη στατιστική και την εξόρυξη δεδομένων (Barga, Roger, et al, 2015).

Η επιχειρησιακή αναλυτική αποτελεί ένα κρίσιμο εργαλείο για τις επιχειρήσεις, καθιστώντας την δυνατή την αναγνώριση, την περιγραφή, και την πρόβλεψη των δραστηριοτήτων τους. Η περιγραφική πλευρά της ανάλυσης επικεντρώνεται στην κατανόηση των τρεχουσών διαδικασιών και πρακτικών. Η προγνωστική διάσταση επιτρέπει την εκτίμηση μελλοντικών τάσεων και εξελίξεων, ενώ η καθοδηγητική αναλυτική υποδεικνύει τις βέλτιστες πρακτικές για βελτιστοποίηση των επιχειρησιακών αποτελεσμάτων. Με αυτό τον τριπλό προσανατολισμό, η επιχειρησιακή αναλυτική είναι ένα ισχυρό εργαλείο που υποστηρίζει τη λήψη αποφάσεων και τη βελτιστοποίηση της αποτελεσματικότητας της επιχείρησης.

2.3 Τεχνολογίες Ψηφιακής Αποθήκευσης Δεδομένων

Η διαδικασία της επιχειρηματικής αναλυτικής αποτελεί αναπόσπαστο μέρος των σύγχρονων τεχνολογικών προσεγγίσεων όπως οι τεχνολογίες ψηφιακής αποθήκευσης δεδομένων. Σε αυτό το πλαίσιο, η ικανότητα αποθήκευσης και ανάκτησης δεδομένων αποτελεί κρίσιμο στοιχείο για την αποτελεσματική λειτουργία της αναλυτικής διαδικασίας. Τα συστήματα αποθήκευσης δεδομένων παρέχουν την υποδομή για τη συλλογή, την οργάνωση και την ανάλυση των δεδομένων, επιτρέποντας στις επιχειρήσεις να αντλούν αδιαλείπτως στρατηγικής σημασίας πληροφορίες για τη λήψη αποφάσεων. Η ικανότητα ανάκτησης δεδομένων σε πραγματικό χρόνο και η αποθήκευση τεράστιων όγκων ποικίλων δεδομένων επιτρέπουν την αποτελεσματική παρακολούθηση των επιδόσεων, την ανίχνευση των τάσεων και την πρόβλεψη μελλοντικών εξελίξεων, προσφέροντας ένα στρατηγικό πλεονέκτημα στον σύγχρονο επιχειρηματικό κόσμο. Επομένως, οι τεχνολογίες ψηφιακής αποθήκευσης δεδομένων αντιπροσωπεύουν τον πυρήνα της συγκρούσεως μεταξύ της αυξανόμενης πολυπλοκότητας των δεδομένων και της ανάγκης για αποτελεσματική διαχείριση και αξιοποίηση τους και αποτελούν την κινητήρια δύναμη πίσω από τη διαδικασία λήψης αποφάσεων, την καινοτομία και την ανάπτυξη. Από τα ευέλικτα Data Lakes, όπου δεδομένα διαφόρων μορφών αποθηκεύονται χωρίς προκαθορισμένο σχήμα, μέχρι τα οργανωμένα

Data Warehouses που επιτρέπουν προηγμένη ανάλυση, οι τεχνολογίες αυτές προσφέρουν ένα φάσμα λύσεων για τη διαχείριση του ραγδαίου όγκου και της ποικιλομορφίας των σύγχρονων δεδομένων.

2.3.1 Data Warehouse

Υπάρχουν δύο διάσημοι ορισμοί της αποθήκης δεδομένων (Data Warehouse) που προέρχονται από τη φυσική τους υλοποίηση (Yessad, Lamia, and Labiod, 2016).

Α) Ο Kimball ορίζει την αποθήκη δεδομένων ως "ένα αντίγραφο δεδομένων συναλλαγών ειδικά δομημένο για αναζήτηση και ανάλυση" (Kimball, Ralph, 1996). Επιπλέον, σύμφωνα με τον Kimball, ο σκοπός μιας αποθήκης δεδομένων είναι "η παροχή πληροφοριών για την υποστήριξη της λήψης αποφάσεων σε μια εταιρεία" (Kimball, Ralph, and Ross, 2013). Επομένως, η αποθήκη δεδομένων είναι μια ειδική βάση δεδομένων που χρησιμοποιείται στο πλαίσιο της λήψης αποφάσεων και της ανάλυσης.

Β) Από την πλευρά του, ο Bill Inmon παρέχει τον ακόλουθο ορισμό: "Μια αποθήκη είναι μια θεματικά προσανατολισμένη (Subject-oriented), ολοκληρωμένη (Integrated), χρονικά μεταβαλλόμενη (Time-variant) και μη πτητική (Non-volatile) συλλογή δεδομένων για την υποστήριξη τη διαδικασία λήψης αποφάσεων της διοίκησης" (INMON, 1996).

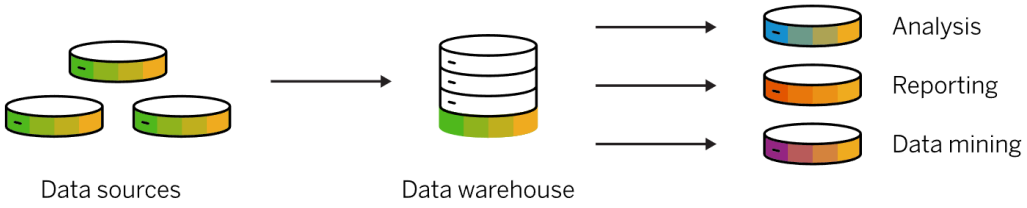
Στη συνέχεια, εξηγούμε τα χαρακτηριστικά που αναφέρονται στον προηγούμενο ορισμό:

- **Προσανατολισμός στο αντικείμενο (Subject-oriented):** τα δεδομένα συνδέονται με την εταιρεία και οργανώνονται ανά λειτουργία.
- **Ολοκληρωμένα (Integrated):** σημαίνει ότι τα δεδομένα που λαμβάνονται από διάφορα επιχειρησιακά και εξωτερικά συστήματα πρέπει να ανταποκρίνονται, τα οποία περιλαμβάνει την επίλυση προβλημάτων λόγω του ορισμού των δεδομένων και διαφορών περιεχομένου, όπως διαφορετικές μορφές και κωδικοποίηση των δεδομένων.
- **Χρονικά μεταβαλλόμενα (Time-variant):** τα δεδομένα προσδιορίζονται ανά συγκεκριμένες περιόδους. Αυτό σημαίνει ότι διατηρούμε το ιστορικό όλων των συναλλαγών.
- **Μη πτητικά (Non-volatile):** τα δεδομένα χρησιμοποιούνται για ερωτήματα και δεν μπορούν να αλλάξουν. Έτσι, οι λειτουργίες ενημέρωσης δεν επιτρέπονται, παρά μόνο η ανάγνωση είναι δυνατή.

Με άλλα λόγια, μια αποθήκη δεδομένων είναι ένα κεντρικό αποθετήριο που αποθηκεύει τα επιχειρησιακά δεδομένα με συγκεκριμένο τρόπο και καθιστά διαθέσιμα και αξιοποιήσιμα για ανάλυση.

Συνοψίζοντας, βασικός σκοπός της αποθήκης δεδομένων (Data Warehouse) είναι να τροφοδοτήσει το business intelligence (BI), τις αναφορές και τα analytics και να υποστηρίξει τις ρυθμιστικές απαιτήσεις, έτσι ώστε οι εταιρείες να μετατρέψουν τα δεδομένα τους σε πληροφορίες και να λάβουν έξυπνες αποφάσεις βάσει δεδομένων. Τα δεδομένα εισάγονται σε μία αποθήκη δεδομένων από λειτουργικά συστήματα όπως το ERP και το CRM, από βάσεις

δεδομένων και από εξωτερικές πηγές όπως τα συστήματα συνεργατών, οι συσκευές Internet of Things (IoT), οι εφαρμογές καιρού και τα μέσα κοινωνικής δικτύωσης.



2.3.2 Data Lake

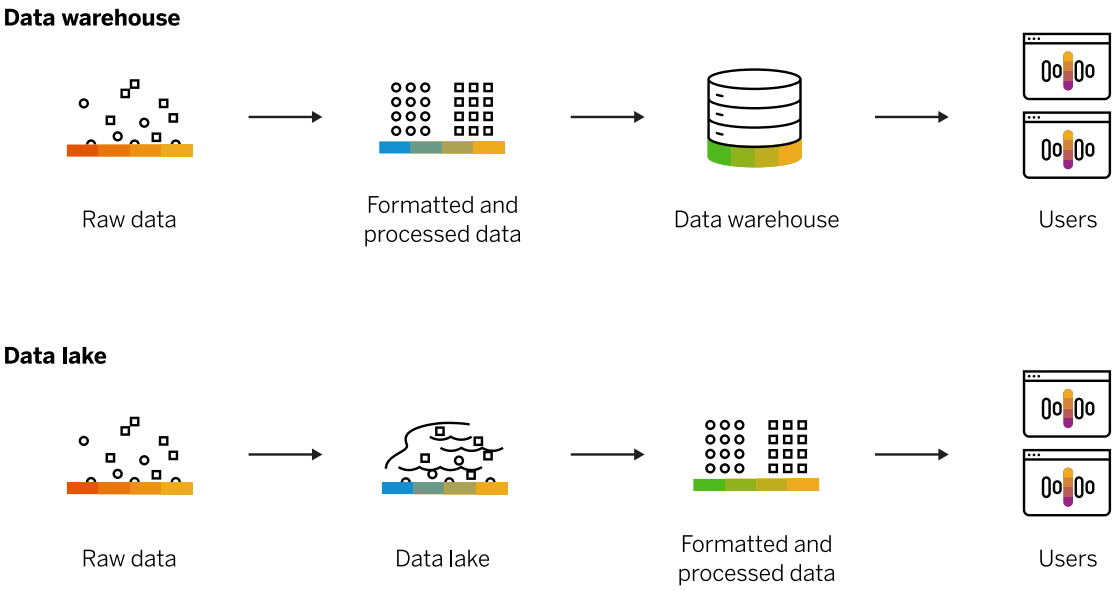
Ένας νέος όρος που ονομάζεται "λίμνη δεδομένων (data lake)" εμφανίστηκε στο προσκήνιο της ψηφιακής εποχής των Big Data. Η απλούστερη έννοια της λίμνης δεδομένων είναι να συγκεντρώσει κάθε δεδομένο οποιασδήποτε μορφής που παράγεται από έναν οργανισμό ή μια επιχείρηση με σκοπό να παρέχει περισσότερες πληροφορίες σε μία πιο λεπτομερή ανάλυση (Khine, Phyu and Wang, 2018). Πιο συγκεκριμένα, μια λίμνη δεδομένων είναι ένα μέρος για την αποθήκευση όλων των ειδών Big Data, όπως δομημένα, μη δομημένα, ημιδομημένα ακόμη και δυαδικά δεδομένα που προέρχονται από επιχειρηματικές εφαρμογές, από εφαρμογές για κινητά, από τα μέσα κοινωνικής δικτύωσης ή από τις συσκευές Internet of Things (IoT). Τα δεδομένα που αποθηκεύονται στη φυσική τους μορφή, συνήθως χρειάζονται περαιτέρω μετατροπή, διαχείριση ή άλλη επεξεργασία ως προς τον τύπο τους, έτσι ώστε να χρησιμοποιηθούν κατάλληλα στην ανάλυση και στην διαχείριση τους. Η πλειοψηφία των data lakes βρίσκεται στο σύννεφο (cloud) λόγω του μεγάλου όγκου δεδομένων που αποθηκεύουν και της ανάγκης για συνδέσεις υψηλής ταχύτητας με κατανεμημένες πηγές. Επίσης, το σύννεφο προσφέρει μεγάλη επεκτασιμότητα στον αποθηκευτικό χώρο με σχετικά μικρό κόστος έναντι των παραδοσιακών βάσεων δεδομένων.

2.3.3 Data Warehouse vs Data Lake

Τόσο η data warehouse όσο και η data lake είναι αποθήκες δεδομένων. Ωστόσο, διαφέρουν σε πολλές πτυχές από τις έννοιες, τις δομές και την υλοποίηση. Οι αποθήκες δεδομένων έχουν σαφώς καθορισμένες ρυθμιστικές λειτουργίες και χωρητικότητα αποθήκευσης. Θεωρητικά, οι λίμνες δεδομένων δεν έχουν κανένα όριο για τη χωρητικότητα αποθήκευσης. Επίσης, οποιοδήποτε είδος δεδομένων με οποιαδήποτε ποσότητα μπορεί να φορτωθεί στο αποθετήριο αποθήκευσης της λίμνης δεδομένων. Οι λίμνες δεδομένων επιτρέπουν στις επιχειρήσεις να βλέπουν πέρα από τον τύπο και τη δομή των δεδομένων, δίνοντάς τους την ευκαιρία να συλλέγουν όσα δεδομένα επιθυμούν. Οι διακρίσεις της λίμνης δεδομένων σε αντίθεση με την επεξεργασία της αποθήκης δεδομένων με πολύ δομημένα δεδομένα, με προ-κατασκευασμένο σχεδιασμό πριν από τη στιγμή της υποβολής ερωτημάτων, με αργά μεταβαλλόμενα δεδομένα είναι οι εξής: (Miloslavskaya, Natalia, and Tolstoy, 2016)

- Ταχεία άφιξη μη δομημένων όγκων δεδομένων.
- Χρήση δυναμικών αναλυτικών εφαρμογών (για ερωτήματα).
- Τα δεδομένα καθίστανται προσβάσιμα αμέσως μετά τη δημιουργία τους, καθώς τα δεδομένα μετασχηματίζονται με βάση λειτουργία του ερωτήματος και τον τομέα της εφαρμογής.

Μία data warehouse αποθηκεύει δεδομένα που μορφοποιήθηκαν για έναν συγκεκριμένο σκοπό, ενώ μία data lake αποθηκεύει δεδομένα στην ανεπεξέργαστη, μη επεξεργασμένη κατάστασή τους. Οι αποθήκες δεδομένων και οι λίμνες συχνά αλληλοσυμπληρώνονται. Για παράδειγμα, όταν τα ανεπεξέργαστα δεδομένα που αποθηκεύονται σε μια λίμνη χρειάζονται για να απαντήσουν σε μια επιχειρηματική ερώτηση (query), μπορούν να εξαχθούν, να εκκαθαριστούν, να μετασχηματιστούν και να χρησιμοποιηθούν σε μια αποθήκη δεδομένων για ανάλυση. Επισημαίνεται πως ο όγκος των δεδομένων, η απόδοση της βάσης δεδομένων και η τιμολόγηση αποθήκευσης παίζουν σημαντικό ρόλο στην επιλογή της καταλληλότερης λύσης αποθήκευσης.



Υποθετικά ένα data lake λειτουργεί σαν ένας λογαριασμός OneDrive, καθώς παρουσιάζει πολλές ομοιότητες. Για παράδειγμα, όταν αποθηκεύουμε ένα σύνολο δεδομένων τοπικά στον υπολογιστή μας τότε έχουμε ορισμένους περιορισμούς, όπως ο όγκος του αποθηκευτικού χώρου, η επεκτασιμότητα και η απομακρυσμένη πρόσβαση. Αντίθετα, όταν τοποθετούμε δεδομένα στον OneDrive λογαριασμό μας οι παραπάνω περιορισμοί εξαλείφονται γιατί πρακτικά αυτά αποθηκεύονται σε ένα data center της Microsoft, το οποίο μπορεί να βρίσκεται οπουδήποτε στον κόσμο , είτε στο απέναντι κτίριο, είτε σε ακτίνα 500 χλμ. μακριά μας. Επομένως, η Microsoft αποθηκεύει τα δεδομένα που τις παρείχαμε για εμάς, καθώς προσφέρει και διαθέτει τον αντίστοιχο εξοπλισμό. Αντίστοιχα, σε επιχειρησιακό επίπεδο το Azure Data Lake εκτελεί την ίδια διαδικασία, δηλαδή συλλέγει όλα τα δεδομένα και τα αποθηκεύει σε μια λίμνη δεδομένων, που την διαχειρίζεται η Microsoft, αντί να τα διατηρεί

ο ίδιος ο οργανισμός, ενδεχομένως σε μηχανήματα που βρίσκονται στις εγκαταστάσεις του. Η Microsoft με την σειρά της εξασφαλίζει την ασφάλεια, την συντήρηση και τα αντίγραφα ασφαλείας αυτών των δεδομένων για λογαριασμό των οργανισμών ή των επιχειρήσεων που την εμπιστεύτηκαν. Επίσης, έχει υποστήριξη ασφαλείας για την λίστα ελέγχου εισόδου, η οποία μας προσφέρει την δυνατότητα να εφαρμόσουμε πολιτικές στην ιεραρχική δομή των αρχείων ώστε όταν συνδεθεί κάποιος να έχει μόνο τα απαραίτητα δικαιώματα και δεδομένα που του αντιστοιχούν για την δουλειά του. Στον παρακάτω πίνακα γίνεται μια σύγκριση μεταξύ της αποθήκης δεδομένων και της λίμνης δεδομένων.

Comparison	Data Warehouse	Data Lake
Data	Structured, processed data	Structured/semi-structured, unstructured data, raw data, unprocessed data
Processing	Schema-on-write	Schema-on-read
Storage	Expensive, reliable	Low-cost storage
Agility	Less agile, fixed configuration	High agility, flexible configuration
Security	Matured	Maturing
Users	Business professional	Data Scientists

2.3.4 Data Lake House

Το data lakehouse είναι ένας τύπος αρχιτεκτονικής δεδομένων που συνδυάζει τα επιθυμητά χαρακτηριστικά ενός data warehousing και μιας data lake, δηλαδή αμβλύνει τις προκλήσεις που παρατηρούνται και στις δύο αυτές τεχνολογίες και αναδεικνύει τα καλύτερα στοιχεία και των δύο (Mazumdar, Dipankar, Hughes, and Onofre, 2023). Ουσιαστικά, μια Lakehouse δεν είναι απλώς μια απλή ενσωμάτωση μεταξύ των δύο τεχνολογιών, αλλά ο σχεδιαστικός της στόχος είναι να συνδυάσει τα πλεονεκτήματα των δύο τεχνολογιών εξαλείφοντας τα μειονεκτήματά τους. (Armbrust, Michael et al, 2021). Πιο επίσημα, μια αρχιτεκτονική λίμνης δεδομένων χαρακτηρίζεται από τα εξής:

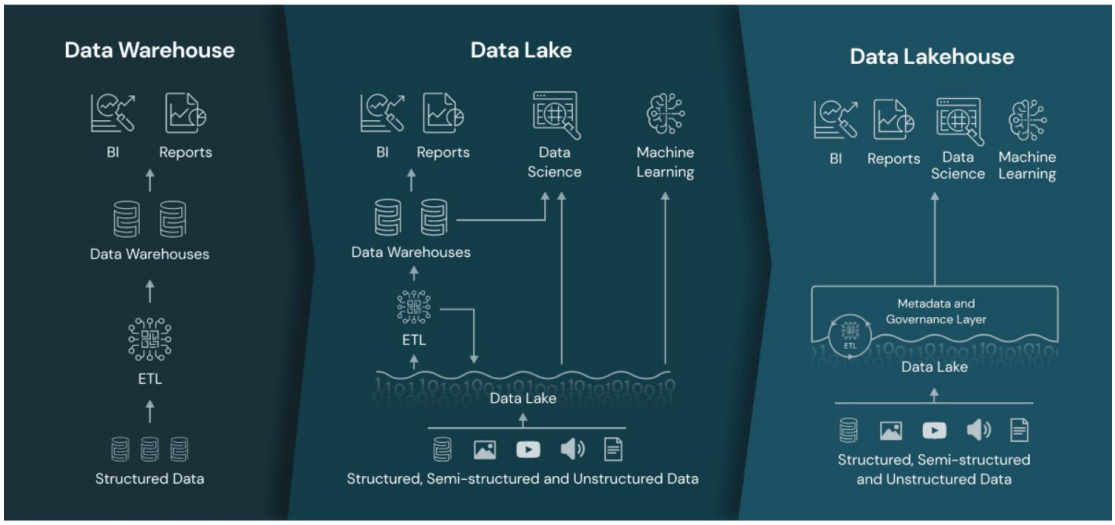
- **Transactional support (Συναλλακτική υποστήριξη)**: Οι data lake houses προσφέρουν αξιοπιστία και συνέπεια (ιδιότητες ACID: Atomicity, Consistency, Isolation, and Durability) σε κάθε συναλλαγή, όπως INSERT ή UPDATE, παρόμοια με ένα data warehouse. Αυτό εξασφαλίζει ταυτόχρονα ασφαλείς αναγνώσεις (read) και εγγραφές (write).
- **Open Data (Ανοιχτά δεδομένα)**: Η βάση ενός data lakehouse είναι η αποθήκευση δεδομένων σε ανοικτές μορφές αρχείων, όπως Apache Parquet, ORC κ.λπ., και μορφές πινάκων, όπως Apache Iceberg [4], Apache Hudi [3] και Delta Lake [6]. Αυτό επιτρέπει την εκτέλεση διαφορετικών αναλυτικών φόρτων εργασίας από διαφορετικές μηχανές, συχνά από διαφορετικούς προμηθευτές ή παρόχους, στα ίδια δεδομένα και αποφεύγεται το κλείδωμα δεδομένων σε μια ιδιόκτητη μορφή.
- **No copy (Δεν υπάρχει αντίγραφο)**: Ένα data lakehouse περιορίζει την αντιγραφή δεδομένων, καθώς ο υπολογιστικός μηχανισμός μπορεί να έχει άμεση πρόσβαση στα

δεδομένα από την πηγή. Το επίπεδο μορφοποίησης πίνακα σε μια αρχιτεκτονική lakehouse προσθέτει ένα λογικό μοντέλο και μια αξιόπιστη διακυβέρνηση στην κορυφή της λίμνης δεδομένων.

- **Data quality and governance (Ποιότητα δεδομένων και διακυβέρνηση)**: Ένα σύστημα lakehouse εστιάζει στη διακυβέρνηση και την ποιότητα των δεδομένων, υιοθετώντας δοκιμασμένες βέλτιστες πρακτικές από τον κόσμο της αποθήκευσης δεδομένων, ώστε να διασφαλίζεται ο κατάλληλος έλεγχος πρόσβασης και να τηρούνται οι κανονιστικές απαιτήσεις.
- **Schema management (Διαχείριση σχημάτων)**: Μια data lakehouse εγγυάται ότι ένα συγκεκριμένο σχήμα τηρείται κατά την εγγραφή νέων δεδομένων στη λίμνη δεδομένων. Διευκολύνει επίσης την εξέλιξη του σχήματος με την πάροδο του χρόνου χωρίς να επιβαρύνεται με το κόστος της επανεγγραφής ολόκληρου του πίνακα.
- **Scalability (Επεκτασιμότητα)**: Μια λίμνη δεδομένων βασίζεται στην ιδέα του διαχωρισμού της αποθήκευσης και του υπολογισμού. Εκμεταλλεύεται τις επιλογές αποθήκευσης χαμηλού κόστους μιας λίμνης δεδομένων, η οποία επιτρέπει την αποθήκευση δεδομένων οποιουδήποτε τύπου (δομημένα, χαλαρά δομημένα κ.λπ.) και όγκου (της τάξης των petabytes) σε ανοικτές μορφές αρχείων όπως το Apache Parquet. Ομοίως, οποιοσδήποτε αναλυτικός φόρτος εργασίας (batch, ad hoc SQL, μηχανική μάθηση) μπορεί να εκτελεστεί στα ανοιχτά αποθηκευμένα δεδομένα και να κλιμακωθεί ανεξάρτητα ανάλογα με τις απαιτήσεις.

Αυτά τα χαρακτηριστικά θέτουν τα θεμελιώδη δομικά στοιχεία για το data lakehousing και επιτρέπουν κάθε είδους δυνατότητες αποθήκευσης δεδομένων σε ένα lakehousing. Τελικά, με μια αρχιτεκτονική lakehouse, ο στόχος είναι να εξαλειφθεί η ανάγκη για μια αρχιτεκτονική δύο επιπέδων για την εκτέλεση διαφορετικών αναλυτικών φόρτων εργασίας και να αποφευχθούν όλες οι πολυπλοκότητες και το κόστος (Mazumdar, Dipankar, Hughes, and Onofre, 2023).

Συγκεκριμένα, ένα data lake house αποθηκεύει τόσο ακατέργαστα, μη επεξεργασμένα δεδομένα όπως μια data lake, όσο και μετασχηματισμένα, δομημένα δεδομένα όπως ένα data warehouse, σε μια ενοποιημένη αρχιτεκτονική. Υπάρχει προαιρετικά η δυνατότητα να αφήσουμε τα δεδομένα εντός του data lake house και να εκτελέσουμε queries σε αυτά, ακόμη και να τα διασφαλίσουμε καλύτερα χωρίς απαραίτητα να τα εξάγουμε. Πολλά από τα οφέλη των σχεσιακών βάσεων δεδομένων (relational databases) υπάρχουν τώρα στο data lake όπως η βελτίωση επιδόσεων και η άμεση εκτέλεση ερωτημάτων (query). Οι λύσεις Data Lake House συνήθως χτίζονται πάνω σε πλατφόρμες αποθήκευσης δεδομένων όπως το Azure Data Lake Storage και χρησιμοποιούν εργαλεία όπως το Apache Spark και το Delta Lake. Επομένως, αυτή η αρχιτεκτονική προσέγγιση επιτρέπει στους οργανισμούς να διατηρούν τα δεδομένα τους σε μια πιο ευέλικτη και ακατέργαστη μορφή ενώ παράλληλα επιτρέπει τη δομημένη επεξεργασία δεδομένων και την βελτίωση απόδοσης ερωτημάτων query. Το Data Lake House είναι ιδιαίτερα σημαντικό σε σύγχρονα σενάρια διαχείρισης δεδομένων που βασίζονται στο cloud.

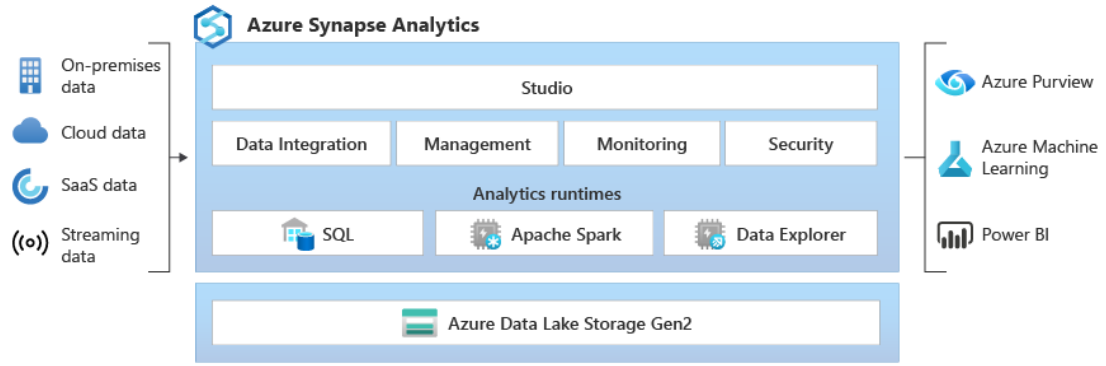


Η ανεξέλεγκτη αύξηση του όγκου και της ποικιλίας των δεδομένων σε συνδυασμό με την ανάγκη για αποτελεσματική αναλυτική επεξεργασία έχει αναδείξει καινοτόμες προσεγγίσεις στον χώρο της διαχείρισης δεδομένων. Οι τεχνολογίες "data lake," "data warehouse," και "data lake house" αποτελούν καίρια στοιχεία σε αυτήν την εξέλιξη. Συνοψίζοντας, το data lake αναφέρεται σε έναν όγκο δεδομένων αποθηκευμένων σε μη δομημένες, ημιδομημένες ή πλήρως δομημένες μορφές, ενώ το data warehouse είναι ένα σύστημα που σχεδιάζεται για την ανάκτηση και ανάλυση μεγάλων όγκων δεδομένων με σκοπό την λήψη αποφάσεων. Ο όρος "data lake house" είναι μια σύγχρονη προσέγγιση που ενσωματώνει τα καλύτερα χαρακτηριστικά των δύο προαναφερθέντων, συνδυάζοντας την ευελιξία του data lake με τη δομική οργάνωση και την αποτελεσματικότητα του data warehouse. Σε αυτό το πλαίσιο, η εισαγωγή του Azure Synapse Analytics αναδεικνύεται ως κορυφαία λύση, επιτρέποντας στις επιχειρήσεις να αντιμετωπίσουν αποτελεσματικά την πρόκληση της διαχείρισης, ανάλυσης και εξαγωγής εργασιών από τα δεδομένα τους, ανοίγοντας νέους ορίζοντες στον κόσμο της επιστήμης των δεδομένων και της επιχειρηματικής αναλυτικής.

2.4 Azure Synapse Analytics

Το Azure Synapse Analytics (ASA) είναι μια σύγχρονη πλατφόρμα ανάλυσης δεδομένων που προσφέρεται ως υπηρεσία και ενσωματώνει την διαχείριση, την αποθήκευση και την ανάλυση δεδομένων έως και το επίπεδο των Big Data. Ουσιαστικά είναι μια υπηρεσία επιχειρησιακής ανάλυσης που επιταχύνει το χρόνο για την εξαγωγή συμπερασμάτων σε αποθήκες δεδομένων και συστήματα μεγάλων δεδομένων. Το Azure Synapse συγκεντρώνει τα καλύτερα στοιχεία των τεχνολογιών SQL που χρησιμοποιούνται στην επιχειρησιακή αποθήκευση δεδομένων, των τεχνολογιών Spark που χρησιμοποιούνται για μεγάλα δεδομένα και του Data Explorer για ανάλυση αρχείων καταγραφής και χρονοσειρών. Επίσης, υποστηρίζει μια ποικιλία εργαλείων, όπως χώρους εργασίας για την ανάπτυξη κώδικα για BI, ML και ELT εντός του Data Lake House. Πιο συγκεκριμένα, προσφέρει υπολογιστικές δυνατότητες χωρίς διακομιστή (Serverless) ή ακόμη και με διακομιστή (Dedicated) για την

ταχύτερη αναζήτηση δεδομένων σε μεγάλη κλίμακα, ανεξαρτήτως όγκου και είδους. Ακόμη, παρέχει δύο διαφορετικούς τύπους υπολογιστικού περιβάλλοντος για διαφορετικούς φόρτους εργασίας. Το ένα είναι το υπολογιστικό περιβάλλον SQL, το οποίο ονομάζεται SQL Pool, και το άλλο είναι το υπολογιστικό περιβάλλον Spark, το οποίο ονομάζεται Spark Pool. Επομένως, οι χρήστες της πλατφόρμας μπορούν να επιλέξουν το υπολογιστικό τους περιβάλλον σύμφωνα με τις επιχειρηματικές τους ανάγκες. Επίσης, το ASA παρέχει στους προγραμματιστές μια ενοποιημένη πύλη που ονομάζεται Synapse Studio, όπου εκεί μπορούν να δημιουργήσουν ένα χώρο εργασίας για την προετοιμασία δεδομένων, τη διαχείριση δεδομένων, την εξερεύνηση δεδομένων, την αποθήκευση δεδομένων, για τα Big Data αλλά και για τις AI εργασίες τους.



Η πλατφόρμα Synapse Analytics υποστηρίζει ικανοποιητικά την απομονωμένη αποθήκευση στο ADLS Gen2 έναντι μιας ποικιλίας επιλογών υπολογισμού, συμπεριλαμβανομένων των Serverless και Dedicated Spark και SQL Pools. Η απομονωμένη αποθήκευση (Isolate Storage) είναι ένας μηχανισμός που παρέχει απομόνωση, ασφάλεια και αποθήκευση δεδομένων, συνδέοντας τον κώδικα με μόνιμα δεδομένα. Η απομονωμένη αποθήκευση έχει σχεδιαστεί για να αποτρέπει τη διαφθορά των δεδομένων και την πρόσβαση σε δεδομένα που αφορούν συγκεκριμένες εφαρμογές, παρέχοντας παράλληλα ένα τυπικό σύστημα αποθήκευσης και ανάκτησης δεδομένων που δεν είναι προσβάσιμο από χρήστες, φακέλους ή εφαρμογές. Επομένως, με τα Serverless Spark και SQL Pools, το ASA είναι σε θέση να κλιμακώνει εύκολα τον υπολογισμό και την αναζήτηση δεδομένων ανεξάρτητα από τον τρόπο αποθήκευσης τους, γεγονός που προσφέρει μεγαλύτερο έλεγχο στο κόστος αποθήκευσης και στο υπολογιστικό κόστος εντός του Data Lake House. Με τα Dedicated SQL Pools, η ASA είναι σε θέση να διατηρήσει την προηγούμενη αποθήκη δεδομένων SQL τύπου MPP εντός του οικοσυστήματός της. Η Synapse SQL είναι ένα κατανεμημένο σύστημα ερωτημάτων για την T-SQL (επέκταση της SQL) που επιτρέπει την αποθήκευση δεδομένων, τα σενάρια ψηφιοποίησης δεδομένων και επεκτείνει την T-SQL για την αντιμετώπιση σεναρίων ροής και μηχανικής μάθησης. [11]

- Η Synapse SQL προσφέρει τόσο μοντέλα χωρίς διακομιστή (Serverless) όσο και μοντέλα αποκλειστικών πόρων (Dedicated). Για προβλέψιμες επιδόσεις και κόστος, δημιουργούνται αποκλειστικές δεξαμενές SQL για την δέσμευση επεξεργαστικής ισχύος για δεδομένα που είναι αποθηκευμένα σε πίνακες SQL. Για απρόσμενους ή δύσκολους φόρτους εργασίας, χρησιμοποιείται το πάντα διαθέσιμο, serverless SQL endpoint.

- Χρήση ενσωματωμένων δυνατοτήτων ροής για την λήψη δεδομένων από πηγές δεδομένων cloud σε πίνακες SQL.
- Ενσωμάτωση της τεχνητής νοημοσύνης στην SQL με τη χρήση μοντέλων μηχανικής μάθησης για τη βαθμολόγηση δεδομένων μέσω της συνάρτησης T-SQL PREDICT.

Το Synapse Analytics ενσωματώνεται και σε άλλες υπηρεσίες της πλατφόρμας δεδομένων Azure για τη καλύτερη διακυβέρνηση δεδομένων, την αποθήκευση, την ασφαλή διαχείριση διαπιστευτηρίων και ταυτότητας, για την υποβολή εκθέσεων και την ανάλυση σε πραγματικό χρόνο.

2.5 Υπολογιστικές Υπηρεσίες

Η εξέλιξη της τεχνολογίας διακρίνεται έντονα στον χώρο των βάσεων δεδομένων, ειδικά με την εμφάνιση δύο κυρίων προσεγγίσεων: των Serverless SQL Pools και των Dedicated SQL Pools. Ο όρος Serverless απεικονίζει μια καινοτόμο προσέγγιση, όπου οι πόροι υποδομής δεν απαιτούν προκαθορισμένη δέσμευση, επιτρέποντας τη δυναμική κλιμάκωση σύμφωνα με τις ανάγκες του χρήστη. Από την άλλη πλευρά, τα Dedicated SQL Pools προσφέρουν αποκλειστικούς πόρους για τις βάσεις δεδομένων, εξασφαλίζοντας υψηλή απόδοση και αξιοπιστία.

2.5.1 Serverless SQL Pools

Η δεξαμενή SQL χωρίς διακομιστή είναι μια υπηρεσία ερωτημάτων στα δεδομένα μιας λίμνης δεδομένων, η οποία επιτρέπει την πρόσβαση σε δεδομένα μέσω των ακόλουθων λειτουργιών:

- Μια οικεία σύνταξη T-SQL για την υποβολή ερωτημάτων σε δεδομένα επί τόπου, χωρίς να χρειάζεται η αντιγραφή ή η φόρτωση δεδομένων σε έναν εξειδικευμένο αποθηκευτικό χώρο.
- Ολοκληρωμένη συνδεσιμότητα μέσω της διεπαφής T-SQL που προσφέρει ένα ευρύ φάσμα εργαλείων επιχειρηματικής ευφυΐας και ad-hoc ερωτημάτων, συμπεριλαμβανομένων των πιο δημοφιλών οδηγών.

Το Serverless SQL Pool είναι ένα κατανεμημένο σύστημα επεξεργασίας δεδομένων, κατασκευασμένο για δεδομένα και υπολογιστικές λειτουργίες μεγάλης κλίμακας. Το Serverless SQL Pool σάς επιτρέπει να αναλύετε τα μεγάλα δεδομένα σας σε δευτερόλεπτα έως λεπτά, ανάλογα με το φόρτο εργασίας. Χάρη στην ενσωματωμένη ανοχή σφαλμάτων εκτέλεσης ερωτημάτων, το σύστημα παρέχει υψηλή αξιοπιστία και υψηλά ποσοστά επιτυχίας ακόμη και για μακροχρόνια ερωτήματα που αφορούν μεγάλα σύνολα δεδομένων. Η δεξαμενή SQL χωρίς διακομιστή είναι χωρίς διακομιστή, επομένως δεν υπάρχει υποδομή για εγκατάσταση ή συστάδες για συντήρηση. Ένα προεπιλεγμένο τελικό σημείο για αυτή την υπηρεσία παρέχεται σε κάθε χώρο εργασίας Azure Synapse, ώστε να μπορείτε να αρχίσετε να ζητάτε δεδομένα αμέσως μετά τη δημιουργία του χώρου εργασίας. Ακόμη, δεν υπάρχει

χρέωση για τη δέσμευση πόρων, χρεώνεστε μόνο για τα δεδομένα που επεξεργάζεστε από τα ερωτήματα που εκτελείτε, επομένως αυτό το μοντέλο είναι ένα πραγματικό μοντέλο πληρωμής ανά χρήση. [9]

Πλεονεκτήματα της δεξαμενής SQL χωρίς διακομιστή

Εάν πρέπει να εξερευνήσετε δεδομένα στη λίμνη δεδομένων, να αποκτήσετε πληροφορίες από αυτά ή να βελτιστοποιήσετε την υπάρχουσα σωλήνωση μετασχηματισμού δεδομένων, μπορείτε να επωφεληθείτε από τη χρήση του serverless SQL pool. Είναι κατάλληλη για τα ακόλουθα σενάρια:

- Βασική ανακάλυψη και εξερεύνηση - Γρήγορη κατανόηση των δεδομένων σε διάφορες μορφές (Parquet, CSV, JSON) στη λίμνη δεδομένων σας, ώστε να μπορείτε να σχεδιάσετε πώς να εξάγετε πληροφορίες από αυτά.
- Λογική αποθήκη δεδομένων - Παρέχετε μια σχεσιακή αφαίρεση πάνω σε ακατέργαστα ή διαφορετικά δεδομένα χωρίς να μετατοπίζετε και να μετασχηματίζετε τα δεδομένα, επιτρέποντας πάντα ενημερωμένη προβολή των δεδομένων σας. Μάθετε περισσότερα για τη δημιουργία λογικής αποθήκης δεδομένων.
- Μετασχηματισμός δεδομένων - Απλός, κλιμακούμενος και αποδοτικός τρόπος μετασχηματισμού δεδομένων στη λίμνη με χρήση T-SQL, ώστε να μπορούν να τροφοδοτηθούν σε BI και άλλα εργαλεία ή να φορτωθούν σε μια σχεσιακή αποθήκη δεδομένων (βάσεις δεδομένων Synapse SQL, Azure SQL Database κ.λπ.).

Κατά την τελευταία δεκαετία έχει διαδοθεί ευρέως η υποδομή υπολογιστικού νέφους (cloud computing infrastructure), όπου οι χρήστες εκκινούν, κατά παραγγελία, εικονικές μηχανές για να αναπτύξουν υπηρεσίες σε μια συστοιχία που τους παρέχεται. Καθώς το υπολογιστικό νέφος συνεχίζει να εξελίσσεται, παρατηρείται μια στροφή προς τη χρήση υπολογιστών χωρίς διακομιστή (serverless computing), όπου η αποθήκευση και ο υπολογισμός διαχωρίζονται τόσο για την παροχή πόρων όσο και για τη χρέωση. Η τάση αυτή ξεκίνησε από υπηρεσίες όπως το Google BigQuery [7] και το AWS Glue [2] που παρέχουν αναλύσεις αποθηκών δεδομένων χωρίς συστοιχίες, ακολουθούμενες από υπηρεσίες όπως το Amazon Athena [1] που επιτρέπουν στους χρήστες να εκτελούν διαδραστικά ερωτήματα σε μια απομακρυσμένη αποθήκη αντικειμένων χωρίς να παρέχουν μια συστάδα υπολογιστών. Ενώ οι προαναφερθείσες υπηρεσίες επικεντρώνονται κυρίως στην παροχή αναλύσεων τύπου SQL, για να καλύψουν την αυξανόμενη ζήτηση, όλοι οι μεγάλοι πάροχοι cloud προσφέρουν πλέον "γενικές" πλατφόρμες υπολογισμού χωρίς διακομιστή, όπως οι AWS Lambda, Google Cloud Functions, Azure Functions και IBM OpenWhisk. Σε αυτές τις πλατφόρμες προγραμματίζονται και εκτελούνται στο σύννεφο βραχύβιες λειτουργίες που ορίζονται από τον χρήστη.

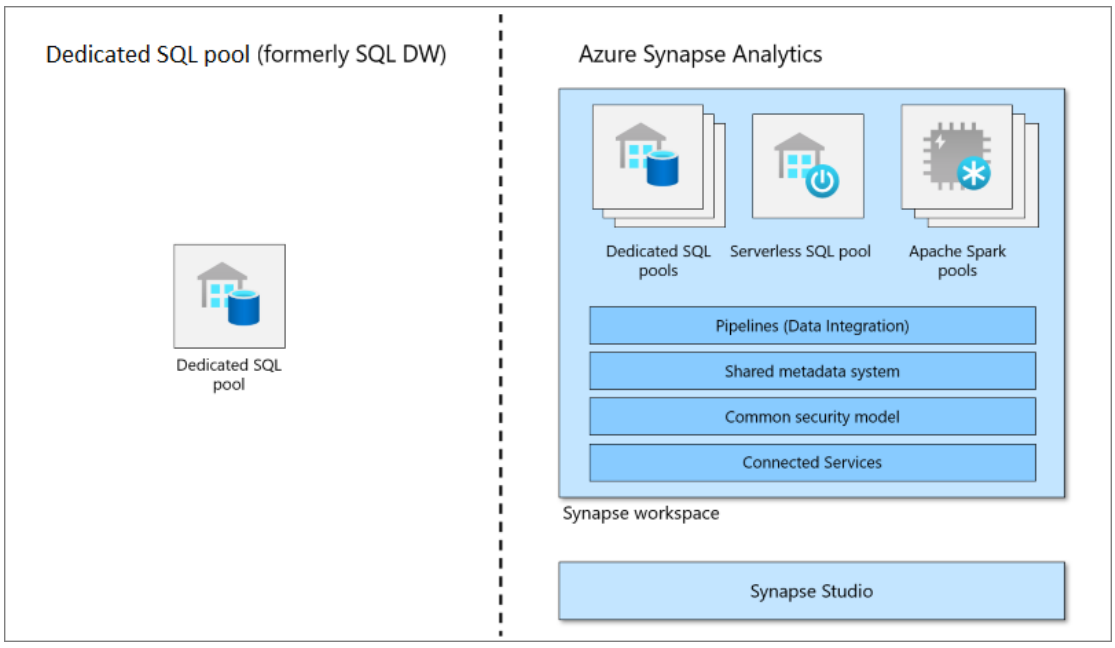
Πρόσφατα, πάροχοι υπηρεσιών νέφους και έργων ανοικτού κώδικα [8], (Hendrickson, Scott, et al, 2016) έχουν προτείνει υπηρεσίες που εκτελούν λειτουργίες στο νέφος ή παρέχουν λειτουργίες ως υπηρεσία. Μέχρι στιγμής, οι λειτουργίες αυτές υπόκεινται σε αυστηρά όρια πόρων. Παρόμοια όρια εφαρμόζουν και άλλοι πάροχοι, όπως οι Google Cloud Functions και Azure Functions. Ανεξάρτητα από αυτούς τους περιορισμούς, οι προσφορές αυτές είναι δημοφιλείς μεταξύ των χρηστών για δύο βασικούς λόγους: την ευκολία ανάπτυξης και την ευέλικτη κατανομή των πόρων. Κατά την ανάπτυξη μιας συστάδας εικονικών μηχανών, οι χρήστες πρέπει να επιλέξουν τον τύπο του instance, τον αριθμό των instances και να

βεβαιωθούν ότι αυτά τα instances τερματίζονται όταν τελειώνει ο υπολογισμός. Αντίθετα, οι serverless υπηρεσίες έχουν ένα πολύ απλούστερο μοντέλο ανάπτυξης, όπου οι λειτουργίες ενεργοποιούνται αυτόματα βάσει συμβάντων, όπως για παράδειγμα η άφιξη νέων δεδομένων.

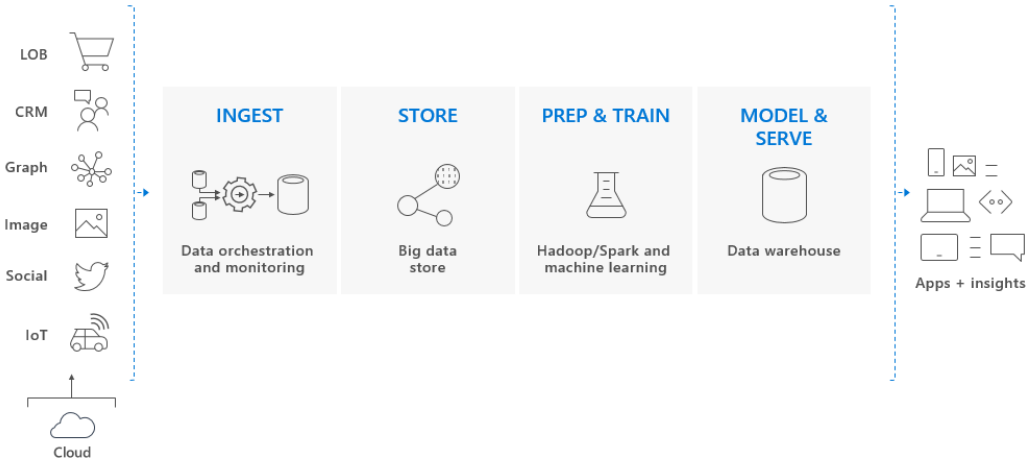
Επομένως, το Serverless SQL Pool [10] επιτρέπει να πραγματοποιούνται ερωτήματα σε αρχεία στους λογαριασμούς μιας υπηρεσίας νέφους, όπως το Azure Storage. Δεν διαθέτει δυνατότητες τοπικής αποθήκευσης ή εισαγωγής. Όλα τα αρχεία στα οποία στοχεύουν τα ερωτήματα είναι εξωτερικά του Serverless SQL Pool. Οτιδήποτε σχετίζεται με την ανάγνωση αρχείων από τον αποθηκευτικό χώρο ενδέχεται να επηρεάσει την απόδοση των ερωτημάτων.

2.5.2 Dedicated SQL Pools

Η SQL Pool είναι η παραδοσιακή αποθήκη δεδομένων (data warehouse). Το Dedicated SQL Pool (πρώην SQL DW) αντιπροσωπεύει μια συλλογή αναλυτικών πόρων που παρέχονται κατά τη χρήση του Synapse SQL. Το μέγεθος μιας Dedicated SQL Pool καθορίζεται από τις μονάδες αποθήκευσης δεδομένων (Data Warehousing Units - DWU). Παλαιότερα ήταν γνωστή ως Azure SQL Data Warehouse, πριν ενταχθεί στην οικογένεια Synapse. Πρόκειται για μια λύση μεγάλων δεδομένων (Big Data) που αποθηκεύει δεδομένα σε μορφή σχεσιακού πίνακα με αποθήκευση σε στήλες. Χρησιμοποιεί επίσης μια αρχιτεκτονική μαζικής παράλληλης επεξεργασίας (MPP) για την αξιοποίηση έως και 60 κόμβων για την εκτέλεση ερωτημάτων. Εφόσον έχουμε τα δεδομένα μας σε ένα Dedicated SQL Pool, μπορούμε να τα αξιοποιήσουμε για ιστορική ανάλυση από ένα ταμπλό, να τα χρησιμοποιήσουμε ως σύνολο δεδομένων για μηχανική μάθηση και για οποιονδήποτε άλλο στόχο δεδομένων που μπορεί να έχουμε για ένα τεράστιο σύνολο δεδομένων. [5]

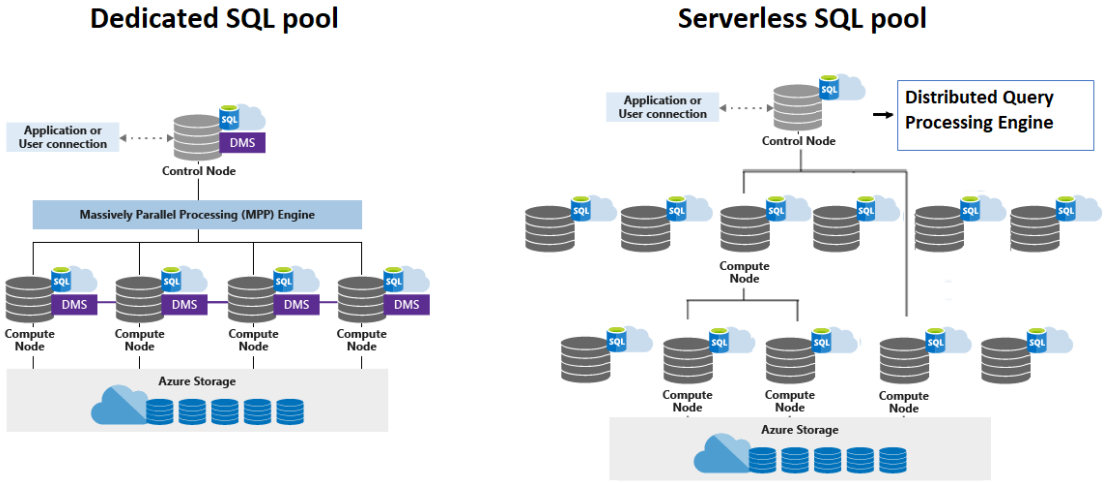


Η αποθήκευση δεδομένων αποτελεί βασικό συστατικό μιας ολοκληρωμένης λύσης μεγάλων δεδομένων που βασίζεται στο cloud. Σε μια λύση δεδομένων νέφους, τα δεδομένα εισάγονται σε αποθήκες μεγάλων δεδομένων από διάφορες πηγές. Μόλις βρεθούν σε ένα μεγάλο κατάστημα δεδομένων, οι αλγόριθμοι Hadoop, Spark και μηχανικής μάθησης προετοιμάζουν και εκπαιδεύουν τα δεδομένα. Όταν τα δεδομένα είναι έτοιμα για σύνθετη ανάλυση, η ειδική δεξαμενή SQL χρησιμοποιεί το PolyBase για να κάνει ερωτήματα στα καταστήματα μεγάλων δεδομένων. Το PolyBase χρησιμοποιεί τυποποιημένα ερωτήματα T-SQL για να φέρει τα δεδομένα σε πίνακες του Dedicated SQL Pool (πρώην SQL DW).



Η αποκλειστική δεξαμενή SQL αποθηκεύει δεδομένα σε σχεσιακούς πίνακες με αποθήκευση σε στήλες. Αυτή η μορφή μειώνει σημαντικά το κόστος αποθήκευσης δεδομένων και βελτιώνει την απόδοση των ερωτημάτων. Μόλις αποθηκευτούν τα δεδομένα, μπορούμε να εκτελέσουμε αναλύσεις σε μαζική κλίμακα. Σε σύγκριση με τα παραδοσιακά συστήματα βάσεων δεδομένων, τα ερωτήματα ανάλυσης ολοκληρώνονται σε δευτερόλεπτα αντί για λεπτά ή σε ώρες αντί για ημέρες. Έτσι, τα αποτελέσματα της ανάλυσης μπορούν να μεταφερθούν σε παγκόσμιες βάσεις δεδομένων ή εφαρμογές αναφοράς. Οι αναλυτές των επιχειρήσεων μπορούν στη συνέχεια να αποκτήσουν πληροφορίες για να λάβουν τεκμηριωμένες επιχειρηματικές αποφάσεις.

Αντίθετα, για τη δεξαμενή SQL χωρίς διακομιστή, η κλιμάκωση γίνεται αυτόματα για να προσαρμόζεται στις απαιτήσεις των πόρων των ερωτημάτων. Καθώς η τοπολογία αλλάζει με την πάροδο του χρόνου με την προσθήκη, την αφαίρεση κόμβων ή τις αποτυχίες, προσαρμόζεται στις αλλαγές και διασφαλίζει ότι το ερώτημά διαθέτει αρκετούς πόρους και ολοκληρώνεται με επιτυχία. Για παράδειγμα, η παρακάτω εικόνα δείχνει πως η Serverless SQL Pool χρησιμοποιεί τέσσερις υπολογιστικούς κόμβους για την εκτέλεση ενός ερωτήματος σε αντίθεση με την αντίστοιχη Dedicated SQL Pool.



2.6 Τεχνητή Νοημοσύνη

2.6.1 Ορισμοί της Τεχνητής Νοημοσύνης

Η Τεχνητή Νοημοσύνη, ή για συντομία ΤΝ, είναι ο τομέας επιστήμης που επιχειρεί να κατανοήσει τη λειτουργία της ανθρώπινης σκέψης και συμπεριφοράς, αλλά και να κατασκευάσει νοήμονες οντότητες, συνδυάζοντας μια τεράστια ποικιλία επιμέρους πεδίων, τα οποία καλύπτουν ένα φάσμα που ξεκινά από γενικούς τομείς, όπως η μάθηση και η αντίληψη και φτάνει σε συγκεκριμένες εργασίες, όπως η απόδειξη μαθηματικών θεωρημάτων και διάγνωση ασθενειών. Συστηματοποιεί και αυτοματοποιεί τις διανοητικές εργασίες και για αυτό μπορεί να έχει εφαρμογή σε οποιαδήποτε σφαίρα της ανθρώπινης διανοητικής δραστηριότητας (Russell & Norvig, 2021). Ένας γενικότερος ορισμός για την τεχνητή νοημοσύνη που θα μπορούσε να είναι ο εξής:

Τεχνητή Νοημοσύνη είναι ο τομέας της επιστήμης των υπολογιστών ο οποίος ασχολείται με τη σχεδίαση και υλοποίηση προγραμμάτων που μπορούν να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που συνήθως αποδίδουμε σε ανθρώπινη συμπεριφορά, όπως η μάθηση, η επίλυση προβλημάτων, η κατανόηση της φυσικής γλώσσας, η επίλυση προβλημάτων κτλ. (Βλαχάβας et al., 2011)

Πέραν αυτού του ορισμού έχουν διατυπωθεί και άλλοι που διαφέρουν σε δύο κύριες διαστάσεις και είναι οργανωμένοι σε τέσσερις κατηγορίες, σύμφωνα με το βιβλίο των Russell & Norvig (2021). Αρχικά, η μία διάσταση τους χωρίζει με βάση το ενδιαφέρον τους ως προς τη συλλογιστική, την διαδικασία με την οποία κάποιος κάνει συλλογισμούς που τον οδηγούν σε κάποια συμπεράσματα, και τις διαδικασίες σκέψης ή ως προς την ενασχόλησή τους με την συμπεριφορά. Η άλλη διάσταση τους χωρίζει ως προς το μέτρο της επιτυχίας με βάση την εγγύτητα προς τις ανθρώπινες επιδόσεις ή σε σχέση με μια ιδανική έννοια νοημοσύνης, η οποία ονομάζεται ορθολογικότητα (rationality). Έτσι υπάρχουν ορισμοί σύμφωνα με τους οποίους ο στόχος της τεχνητής νοημοσύνης είναι να φτιάξει:

1. Συστήματα που σκέφτονται σαν τον άνθρωπο
    - "Η συναρπαστική νέα προσπάθεια για να κάνουμε τους υπολογιστές να σκέφτονται… μηχανές με νόηση, με την πλήρη και κυριολεκτική έννοια." (Haugeland, 1985)
    - "Η αυτοματοποίηση των δραστηριοτήτων που συσχετίζουμε με την ανθρώπινη σκέψη, όπως η λήψη αποφάσεων, η επίλυση προβλημάτων, η μάθηση..." (Bellman, 1978)
2. Συστήματα που σκέφτονται ορθολογικά
    - "Η μελέτη των νοητικών ικανοτήτων με τη χρήση υπολογιστικών μοντέλων." (Charniak & McDermott, 1985)
    - "Η μελέτη των υπολογιστικών εργασιών που μας δίνουν τη δυνατότητα να αντιλαμβανόμαστε, να συλλογιζόμαστε και να ενεργούμε." (Winston, 1992)
3. Συστήματα που ενεργούν σαν τον άνθρωπο
    - "Η τέχνη της δημιουργίας μηχανών που πραγματοποιούν λειτουργίες οι οποίες απαιτούν νοημοσύνη όταν πραγματοποιούνται από ανθρώπους." (Kurzweil, 1990)
    - "Η μελέτη του πώς μπορούμε να κάνουμε τους υπολογιστές να κάνουν πράγματα στα οποία, προς το παρόν, οι άνθρωποι είναι καλύτεροι." (Rich & Knight, 1991)
4. Συστήματα που ενεργούν ορθολογικά
    - "Υπολογιστική νοημοσύνη είναι η μελέτη της σχεδίασης ευφυών πρακτόρων." (Poole et al., 1998)
    - "Η τεχνητή νοημοσύνη ασχολείται με την ευφυή συμπεριφορά των τεχνουργημάτων." (Nilsson, 1998).

Ιστορικά, έχουν ακολουθηθεί και οι τέσσερις παραπάνω προσεγγίσεις στην τεχνητή νοημοσύνη. Αναπόφευκτα υπάρχει κάποια διένεξη σε αυτές που εστιάζουν στον άνθρωπο και σε αυτές που εστιάζουν στην ορθολογικότητα, καθώς όταν κάνουμε διάκριση μεταξύ της ανθρώπινης και της ορθολογικής συμπεριφοράς δεν υπονοούμε αναγκαστικά ότι οι άνθρωποι είναι "παράλογοι" με την έννοια της συναισθηματικής αστάθειας ή της παραφροσύνης. Επομένως, μια ανθρωποκεντρική προσέγγιση θα πρέπει να είναι εμπειρική επιστήμη με υποθέσεις και πειραματική επιβεβαίωση, ενώ μια ορθολογιστική προσέγγιση να περιλαμβάνει ένα συνδυασμό μαθηματικών και μηχανικής.

2.6.2 Ιστορική Αναδρομή

Η ιστορία της τεχνητής νοημοσύνης (ΤΝ) είναι ένα συναρπαστικό ταξίδι που εκτείνεται σε αρκετές δεκαετίες και σημαδεύεται από σημαντικά ορόσημα, ανακαλύψεις και αλλαγές παραδείγματος στην τεχνολογία και τη σκέψη. Οι ρίζες της τεχνητής νοημοσύνης μπορούν να εντοπιστούν στην αρχαιότητα, όπου οι μύθοι και οι ιστορίες απεικόνιζαν ευφυείς, ανθρωποειδείς μηχανές. Οι απαρχές της τεχνητής νοημοσύνης ανάγονται στους "συλλογισμούς" του Αριστοτέλη (384-322 π.Χ.), έναν τρόπο κωδικοποίησης της ορθής σκέψης μέσω διαφόρων κανόνων που ανάλυαν τη διαδικασία της σκέψης και αποτέλεσαν τη

βάση του πεδίου της λογικής (Russell & Norvig, 2021, Βλαχάβας et al., 2011). Ωστόσο, η επίσημη έναρξη της τεχνητής νοημοσύνης ως τομέα μελέτης έγινε στα μέσα του 20ού αιώνα.

- Δεκαετία 1950-1960: Η γέννηση της ΤΝ και της Συμβολικής Λογικής

Ο όρος "τεχνητή νοημοσύνη" επινοήθηκε από τον επιστήμονα πληροφορικής John McCarthy το 1956 κατά τη διάρκεια του συνεδρίου του Dartmouth, το οποίο θεωρείται η γέννηση της ΤΝ ως επιστημονικού κλάδου. Η πρώιμη έρευνα για την τεχνητή νοημοσύνη επικεντρώθηκε στον συμβολικό συλλογισμό και τη λογική, με την πεποίθηση ότι η ανθρώπινη νοημοσύνη θα μπορούσε να αναπαραχθεί μέσω συστημάτων βασισμένων σε κανόνες. Οι ερευνητές ανέπτυξαν προγράμματα ικανά να επιλύουν μαθηματικά προβλήματα, να παίζουν παιχνίδια όπως το σκάκι και η ντάμα, ακόμα και να προσομοιώνουν την κατανόηση της φυσικής γλώσσας.

- Δεκαετία 1970-1980: Αναπαράσταση γνώσης και συστήματα εμπειρογνωμόνων

Κατά τις δεκαετίες του 1970 και 1980, η έρευνα της ΤΝ μετατοπίστηκε προς την αναπαράσταση γνώσης και τα συστήματα εμπειρογνωμόνων. Οι ερευνητές στόχευαν στην κωδικοποίηση της ανθρώπινης εμπειρογνωμοσύνης σε συστήματα υπολογιστών για την επίλυση πολύπλοκων προβλημάτων. Κατά τη διάρκεια αυτής της περιόδου αναπτύχθηκαν συστήματα εμπειρογνωμόνων, όπως το MYCIN για ιατρική διάγνωση και το DENDRAL για χημική ανάλυση. Ωστόσο, τα συστήματα αυτά είχαν περιορισμούς, καθώς στηρίζονταν σε μεγάλο βαθμό σε ρητά προγραμματισμένους κανόνες και αντιμετώπιζαν προβλήματα αβεβαιότητας.

- Δεκαετία 1980-1990: Χειμώνας της ΤΝ και συνδεσιμότητα

Στα τέλη της δεκαετίας του 1980 παρατηρήθηκε πτώση της έρευνας ΤΝ, γνωστή ως "χειμώνας της ΤΝ", λόγω ανεκπλήρωτων προσδοκιών και υπερβολικών υποσχέσεων. Η χρηματοδότηση για έργα τεχνητής νοημοσύνης αλλά και το ενδιαφέρον μειώθηκε. Ωστόσο, κατά τη διάρκεια αυτής της περιόδου εμφανίστηκε η συνδεσιμότητα ως μια εναλλακτική προσέγγιση της συμβολικής συλλογιστικής. Τα νευρωνικά δίκτυα, εμπνευσμένα από τη δομή του ανθρώπινου εγκεφάλου, κέρδισαν την προσοχή. Αν και η πρόοδος ήταν αργή, ερευνητές όπως ο Geoffrey Hinton συνέβαλαν στην ανάπτυξη αλγορίθμων αντίστροφης διάδοσης, αναζωπυρώνοντας το ενδιαφέρον για τα νευρωνικά δίκτυα.

- Δεκαετία 1990-2000: Μηχανική Μάθηση και πρακτικές εφαρμογές

Οι εξελίξεις στη μηχανική μάθηση, ιδίως σε αλγορίθμους όπως οι μηχανές διανυσμάτων υποστήριξης και τα δέντρα αποφάσεων, σημάδεψαν τη δεκαετία του 1990 και τις αρχές της δεκαετίας του 2000. Η διαθεσιμότητα μεγάλων συνόλων δεδομένων και η αυξημένη υπολογιστική ισχύς τροφοδότησαν την επανεμφάνιση της ΤΝ. Πρακτικές εφαρμογές εμφανίστηκαν σε τομείς όπως η αναγνώριση ομιλίας, η όραση υπολογιστών και η επεξεργασία φυσικής γλώσσας. Η νίκη του Deep Blue της IBM επί του πρωταθλητή σκακιού Garry Kasparov το 1997 και η ανάπτυξη του διαδικτύου συνέβαλαν στην ανανέωση της αισιοδοξίας για την τεχνητή νοημοσύνη.

- Δεκαετία 2010-2020: Βαθιά μάθηση, μεγάλα δεδομένα και ΤΝ παντού

Την τελευταία δεκαετία παρατηρήθηκε η κυριαρχία της βαθιάς μάθησης, μιας υποκατηγορίας της μηχανικής μάθησης που χρησιμοποιεί νευρωνικά δίκτυα με πολλαπλά επίπεδα. Τα επιτεύγματα στη βαθιά μάθηση, που τροφοδοτήθηκαν από τα μεγάλα δεδομένα και τις ισχυρές GPU, οδήγησαν σε αξιοσημείωτα επιτεύγματα στην αναγνώριση εικόνων, τη σύνθεση ομιλίας και την κατανόηση της φυσικής γλώσσας. Τεχνολογίες όπως η Siri, η Alexa και τα αυτοκινούμενα αυτοκίνητα έγιναν μέρος της καθημερινότητάς μας. Η τεχνητή νοημοσύνη διαδραμάτισε επίσης κρίσιμο ρόλο σε τομείς όπως η υγειονομική περίθαλψη, η οικονομία και η κυβερνοασφάλεια.

- Σήμερα: Προκλήσεις και μελλοντικές προοπτικές

Ενώ η τεχνητή νοημοσύνη έχει κάνει τεράστια βήματα προόδου, οι προκλήσεις παραμένουν, συμπεριλαμβανομένων των ηθικών ανησυχιών, της προκατάληψης των αλγορίθμων και της ανάγκης για εξηγήσιμη τεχνητή νοημοσύνη. Η τρέχουσα έρευνα επικεντρώνεται στη δημιουργία πιο ισχυρών, διαφανών και υπεύθυνων συστημάτων ΤΝ. Το μέλλον της τεχνητής νοημοσύνης υπόσχεται εξελίξεις σε τομείς όπως η κβαντική υπολογιστική, η ενισχυτική μάθηση και η αυτοματοποίηση με βάση την τεχνητή νοημοσύνη, διαμορφώνοντας έναν κόσμο όπου τα ευφυή συστήματα θα συνεργάζονται απρόσκοπτα με τον άνθρωπο.

2.6.3 Δοκιμασία Turing

Η δοκιμή Turing, που προτάθηκε από τον Βρετανό μαθηματικό και επιστήμονα υπολογιστών Alan Turing το 1950, είναι μια θεμελιώδης έννοια στον τομέα της τεχνητής νοημοσύνης (ΤΝ) και εξακολουθεί να αποτελεί σημείο αναφοράς για την αξιολόγηση της νοημοσύνης των μηχανών. Στόχος του Turing ήταν να απαντήσει στο θεμελιώδες ερώτημα: Μπορούν οι μηχανές να επιδείξουν νοημοσύνη που μοιάζει με την ανθρώπινη; Για να θέσει σε λειτουργία αυτό το ερώτημα, επινόησε μια δοκιμασία που θα αξιολογούσε την ικανότητα μιας μηχανής να συμμετέχει σε συνομιλία σε φυσική γλώσσα σε σημείο που να μην διακρίνεται από έναν άνθρωπο. Η ουσία του test Turing έγκειται σε ένα παιχνίδι μίμησης. Στην αρχική του μορφή, ο Turing οραματίστηκε ένα σενάριο όπου ένας ανακριτής αλληλεπιδρά τόσο με έναν άνθρωπο όσο και με μια μηχανή, χωρίς να γνωρίζει ποιος είναι ποιος. Ο ρόλος του ανακριτή είναι να διακρίνει ποιος από τους συμμετέχοντες είναι η μηχανή, βασιζόμενος αποκλειστικά στις απαντήσεις τους σε ερωτήσεις ή προτροπές. Εάν ο ανακριτής δεν μπορεί να διακρίνει αξιόπιστα μεταξύ του ανθρώπου και της μηχανής, τότε λέγεται ότι η μηχανή έχει περάσει τη δοκιμή Turing. Ο Turing άφησε σκόπιμα ασαφή τα κριτήρια της "νοημοσύνης" και της "συνομιλίας που μοιάζει με ανθρώπινη", καθώς πίστευε ότι ο ακριβής ορισμός αυτών των όρων θα ήταν προβληματικός. Αντ' αυτού, επικεντρώθηκε στο πρακτικό, παρατηρήσιμο αποτέλεσμα της συνομιλίας. Εάν μια μηχανή μπορούσε να μιμηθεί την ανθρώπινη συνομιλία αρκετά καλά ώστε να διαφεύγει της ανίχνευσης, τότε επέδειξε μια μορφή νοημοσύνης, τουλάχιστον στο πλαίσιο της γλωσσικής αλληλεπίδρασης.

Η δοκιμή Turing διαδραμάτισε καθοριστικό ρόλο στη διαμόρφωση των συζητήσεων σχετικά με την τεχνητή νοημοσύνη και τη νοημοσύνη των μηχανών. Έχει προκαλέσει συζητήσεις σχετικά με τη φύση της συνείδησης, τα όρια της τεχνητής νοημοσύνης και τις ηθικές επιπτώσεις της δημιουργίας μηχανών που μπορούν να μιμηθούν πειστικά την ανθρώπινη

συμπεριφορά. Αν και το να περάσει το Turing Test δεν σημαίνει απαραίτητα ότι μια μηχανή διαθέτει πραγματική κατανόηση ή συνείδηση, χρησιμεύει ως ένα ρεαλιστικό σημείο αναφοράς για την αξιολόγηση της αποτελεσματικότητας των συστημάτων τεχνητής νοημοσύνης στην επικοινωνία που μοιάζει με την ανθρώπινη. Με την πάροδο των ετών, διάφορα συστήματα τεχνητής νοημοσύνης έχουν υποβληθεί στη δοκιμασία Turing, με ορισμένα να έχουν σημειώσει αξιοσημείωτη επιτυχία σε συγκεκριμένους τομείς. Ωστόσο, το τεστ αντιμετώπισε και επικρίσεις. Ορισμένοι υποστηρίζουν ότι θέτει χαμηλό πήχη για τη νοημοσύνη, καθώς εστιάζει κυρίως σε συμπεριφορές επιφανειακού επιπέδου και όχι σε βαθύτερες γνωστικές ικανότητες. Άλλοι τονίζουν τη σημασία της αξιολόγησης της νοημοσύνης μέσω ενός ευρύτερου φάσματος εργασιών και πλαισίων.

Τέλος, καθώς η Τεχνητή Νοημοσύνη συνεχίζει να εξελίσσεται, το Turing Test παραμένει μια θεμελιώδης έννοια, αλλά οι ερευνητές διερευνούν πιο διαφοροποιημένες και ολοκληρωμένες προσεγγίσεις για την αξιολόγηση της νοημοσύνης των μηχανών. Αυτές περιλαμβάνουν την ενσωμάτωση στοιχείων της λογικής της κοινής λογικής, της συναισθηματικής νοημοσύνης και της ικανότητας χειρισμού διφορούμενων ή πρωτότυπων καταστάσεων -παράγοντες που υπερβαίνουν το πεδίο εφαρμογής της αρχικής διατύπωσης του Turing. Παρά τους περιορισμούς της, η δοκιμασία Turing αποτελεί ορόσημο στη φιλοσοφική και πρακτική διερεύνηση της νοημοσύνης των μηχανών και στην προσπάθειά μας να κατανοήσουμε τα όρια μεταξύ της ανθρώπινης και της τεχνητής νόησης.

2.6.4 Πεδία εφαρμογής

Στην βιβλιογραφία μπορούμε να εντοπίσουμε πλειάδα πεδίων εφαρμογής της τεχνητής νοημοσύνης, σε αυτό το κομμάτι θα επικεντρωθούμε στα σημαντικότερα και ευρύτερα πεδία έτσι όπως κατηγοριοποιούνται από την Pannu (2015) , ώστε να μπορέσουμε να καλύψουμε το μέγιστο δυνατό πλήθος:

Α. Γλωσσική κατανόηση: Η ικανότητα «κατανόησης» και ανταπόκρισης στη φυσική γλώσσα. Να μεταφράζει από προφορική γλώσσα σε γραπτή μορφή και να μεταφράζει από μια φυσική γλώσσα σε μια άλλη φυσική γλώσσα.

- Κατανόηση ομιλίας
- Επεξεργασία σημασιολογικής πληροφορίας
- Ερώτηση και απάντηση
- Ανάκτηση πληροφοριών
- Μετάφραση γλώσσας

Β. Επίλυση προβλημάτων: Ικανότητα διατύπωσης ενός προβλήματος σε κατάλληλη αναπαράσταση, προγραμματισμό για την επίλυσή του και γνώσης πότε χρειάζονται νέες πληροφορίες και πώς να τις αποκτήσει.

- Συμπέρασμα (απόδειξη θεωρήματος βασισμένο σε ανάλυση, λογικό επαγωγικό συμπέρασμα)
- Διαδραστική επίλυση προβλήματος
- Αυτόματη εγγραφή προγράμματος

- Ευρετική αναζήτηση

C. Συστήματα μάθησης και προσαρμογής: Η ικανότητα προσαρμογής της συμπεριφοράς με βάση την προηγούμενη εμπειρία και η ανάπτυξη γενικών κανόνων που αφορούν τον κόσμο με βάση αυτή την εμπειρία.

- Κυβερνητική, βασική έννοια της κυβερνητικής είναι η ανατροφοδότηση, όπου τα παρατηρούμενα αποτελέσματα των ενεργειών λαμβάνονται ως εισροές για περαιτέρω δράση με τρόπους που υποστηρίζουν την επιδίωξη και τη διατήρηση συγκεκριμένων συνθηκών ή τη διακοπή τους
- Σχηματισμός έννοιας, η διαδικασία ταξινόμησης των παραδειγμάτων σε τάξεις με νόημα

D. Αντίληψη: Η ικανότητα ανάλυσης μιας σκηνής συνδέοντάς την με ένα εσωτερικό μοντέλο που αντιπροσωπεύει τη «γνώση του κόσμου» του οργανισμού που αντιλαμβάνεται. Το αποτέλεσμα αυτής της ανάλυσης είναι ένα δομημένο σύνολο σχέσεων μεταξύ οντοτήτων.

- Αναγνώριση μοτίβου
- Ανάλυση Σκηνής

E. Μοντελοποίηση: Η ικανότητα ανάπτυξης μιας εσωτερικής αναπαράστασης και ενός συνόλου κανόνων μετασχηματισμού που μπορούν να χρησιμοποιηθούν για την πρόβλεψη της συμπεριφοράς και της σχέσης μεταξύ κάποιου συνόλου αντικειμένων ή οντοτήτων του πραγματικού κόσμου.

- Το πρόβλημα αναπαράστασης για συστήματα επίλυσης προβλημάτων
- Μοντελοποίηση Φυσικών Συστημάτων (Οικονομικά, Κοινωνιολογικά, Οικολογικά κ.λπ.)

F. Ρομπότ: Συνδυασμός των περισσότερων ή όλων των παραπάνω ικανοτήτων με την ικανότητα κίνησης σε έδαφος και χειρισμού αντικειμένων.

- Εξερεύνηση
- Μεταφορές/Πλοήγηση
- Βιομηχανικός αυτοματισμός (π.χ. Έλεγχος Διαδικασιών, Εργασίες Συναρμολόγησης, Εκτελεστικές εργασίες)
- Ασφάλεια
- Στρατός
- Νοικοκυριό
- Άλλα (Γεωργία, Αλιεία, Μεταλλεία, Υγιεινή, Κατασκευές κ.λπ.)

G. Παιχνίδια: Η ικανότητα αποδοχής ενός επίσημου συνόλου κανόνων για παιχνίδια όπως Chess, Go, Kalah, Checkers, κ.λπ., και να μεταφραστούν αυτοί οι κανόνες σε μια αναπαράσταση ή δομή που επιτρέπει τη χρήση των ικανοτήτων επίλυσης προβλημάτων και μάθησης για την επίτευξη ενός επαρκές επίπεδο απόδοσης.

2.7 Μηχανική Μάθηση

2.7.1 Ορισμός Μηχανικής Μάθησης

Η μηχανική μάθηση (Machine Learning - ML) είναι ένα υποσύνολο της τεχνητής νοημοσύνης (AI) και κατ' επέκταση της επιστήμης των υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση μπορεί να οριστεί ως το φαινόμενο κατά το οποίο ένα σύστημα βελτιώνει την απόδοσή του κατά την εκτέλεση μιας συγκεκριμένης εργασίας, χωρίς να υπάρχει ανάγκη να προγραμματιστεί εκ νέου (Γεωργούλη, 2015). Συγκεκριμένα, η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα παρεχόμενα δεδομένα και έπειτα να κάνουν προβλέψεις σχετικά με αυτά. Ουσιαστικά, επικεντρώνεται στη διαδικασία αυτόματης διδασκαλίας των υπολογιστών και βελτίωσης των αποτελεσμάτων, χωρίς την χρήση εξωγενούς προγραμματισμού ή άλλων μέσων.

Ένας σχετικός γενικός ορισμός Μηχανικής Μάθησης δίνεται από τον Mitchell (1997): «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E.»

Στη μηχανική μάθηση, οι αλγόριθμοι εκπαιδεύονται για να βρουν μοτίβα και συσχετίσεις σε μεγάλα σύνολα δεδομένων και να λαμβάνουν τις καλύτερες αποφάσεις και προβλέψεις με βάση αυτή την ανάλυση. Οι εφαρμογές μηχανικής μάθησης βελτιώνονται επαναληπτικά καθώς συναντούν νέα δεδομένα, επιτρέποντας στο σύστημα να εξελίσσεται συνεχώς και να βελτιώνει τις επιδόσεις του. Η συμβατική μέθοδος προγραμματισμού αποτελείται από δύο διακριτά βήματα. Δεδομένης μιας προδιαγραφής για το πρόγραμμα, δηλαδή, τι υποτίθεται ότι πρέπει να κάνει το πρόγραμμα και όχι πώς, το πρώτο βήμα είναι να δημιουργήσουμε μια λεπτομερή σχεδίαση για το πρόγραμμα, δηλαδή ένα σταθερό σύνολο βημάτων ή κανόνων για την επίλυση του προβλήματος. Δεύτερο βήμα είναι η εφαρμογή του λεπτομερούς σχεδιασμού ως πρόγραμμα σε γλώσσα υπολογιστή (Rebala et al., 2019). Οι αλγόριθμοι της Μηχανικής Μάθησης μπορούν να λύσουν πολλά από τα δύσκολα προβλήματα με γενικό τρόπο. Αυτοί οι αλγόριθμοι δεν απαιτούν σαφή λεπτομερή σχεδιασμό. Αντίθετα, μαθαίνουν ουσιαστικά τη λεπτομερή σχεδίαση από ένα σύνολο δεδομένων, δηλαδή, ένα σύνολο παραδειγμάτων που απεικονίζουν τη συμπεριφορά του προγράμματος. Με άλλα λόγια, μαθαίνουν από δεδομένα. Όσο μεγαλύτερο είναι το σύνολο δεδομένων, τόσο πιο ακριβή γίνονται. (Rebala et al., 2019) Ο στόχος ενός αλγορίθμου Μηχανικής Μάθησης είναι να μάθει ένα μοντέλο ή ένα σύνολο κανόνων από ένα σύνολο δεδομένων, ώστε να μπορεί να προβλέψει σωστά τα στοιχεία των δεδομένων που δεν βρίσκονται στο σύνολο δεδομένων. Οι αλγόριθμοι της Μηχανικής Μάθησης τείνουν να είναι πιο ακριβείς από τους κανόνες που έχουν δημιουργηθεί από τον άνθρωπο, καθώς λαμβάνουν υπόψη όλα τα στοιχεία των δεδομένων σε ένα σύνολο δεδομένων χωρίς καμία ανθρώπινη προκατάληψη λόγω προηγούμενης γνώσης.

2.7.2 Τύποι Μηχανικής Μάθησης

Η μηχανική μάθηση περιλαμβάνει διάφορες προσεγγίσεις, κάθε μία από τις οποίες είναι κατάλληλη για διαφορετικούς τύπους εργασιών και δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης ταξινομούνται με βάση το επιθυμητό αποτέλεσμα του αλγορίθμου. Οι κύριοι τύποι μηχανικής μάθησης είναι οι εξής:

- Εποπτευόμενη μάθηση ή μάθηση με επίβλεψη (Supervised Learning)
- Μη Εποπτευόμενη μάθηση ή μάθηση χωρίς επίβλεψη (Unsupervised Learning)
- Ημι-εποπτευόμενη μάθηση (Semi-supervised Learning)
- Ενισχυτική Μάθηση (Reinforcement Learning)

1. **Εποπτευόμενη μάθηση (Supervised Learning)**: Η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο (Γεωργούλη, 2015). Με απλά λόγια ο αλγόριθμος μαθαίνει να αντιστοιχίζει τις εισόδους στις εξόδους, κάνοντας προβλέψεις ή ταξινομήσεις σε νέα, αθέατα δεδομένα. Αυτή η προσέγγιση χρησιμοποιείται ευρέως σε εργασίες όπως η αναγνώριση εικόνων, η αναγνώριση ομιλίας και η επεξεργασία φυσικής γλώσσας Η εποπτευόμενη μάθηση εμπίπτει κυρίως σε δύο κατηγορίες: (Rebala et al., 2019)

- Ταξινόμηση (Classification): Τα προβλήματα ταξινόμησης αναφέρονται στην ικανότητα να ταξινομηθεί κάτι σε ένα ξεχωριστό σύνολο κλάσεων ή κατηγοριών. Η αναγνώριση ενός αντικειμένου σε σκύλο, γάτα, αεροπλάνο κ.λπ. είναι ένα παράδειγμα αναγνώρισης της κατηγορίας του αντικειμένου. Το αν το χρηματιστήριο πρόκειται να δει σημαντική αύξηση ή μείωση ή καμία σημαντική αλλαγή είναι ένα άλλο παράδειγμα αυτής της κατηγορίας. Εδώ οι τρεις κατηγορίες είναι αύξηση, μείωση και καμία αλλαγή.
- Παλινδρόμηση (Regression): Η παλινδρόμηση αναφέρεται στην ικανότητα πρόβλεψης τιμών μιας συνεχούς μεταβλητής, για παράδειγμα, ενός μοντέλου για την πρόβλεψη της τιμής της μετοχής σε καθημερινή ή εβδομαδιαία βάση.

2. **Μη εποπτευόμενη μάθηση (Unsupervised Learning)** αποτελεί κατηγορία της μηχανικής μάθησης, στόχος της οποίας είναι η ανακάλυψη πιθανής δομής που μπορεί να κρύβεται πίσω από μη χαρακτηρισμένα δεδομένα. Ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους (Γεωργούλη, 2015). Με απλά λόγια, ο αλγόριθμος εντοπίζει μοτίβα και σχέσεις χωρίς ρητή καθοδήγηση. Η ομαδοποίηση και η μείωση των διαστάσεων είναι κοινές εφαρμογές της μάθησης χωρίς επίβλεψη, βοηθώντας στην αποκάλυψη κρυφών δομών μέσα στα δεδομένα και στην απλοποίηση της αναπαράστασής τους. Ο αλγόριθμος θα αναγνωρίσει συμπλέγματα ή ομάδες παρόμοιων στοιχείων ή ομοιότητα νέου στοιχείου με υπάρχουσα ομάδα κλπ. (Rebala et al., 2019). Χρησιμοποιείται κυρίως σε προβλήματα:

- Ανάλυσης Συσχετισμών (Association Analysis)
- Ομαδοποίησης (Clustering)

3. **Ημι-εποπτευόμενη μάθηση (Semi-supervised Learning).** Η ημι-εποπτευόμενη μάθηση εμπίπτει κάπου μεταξύ εποπτευόμενης και μη εποπτευόμενης μάθησης. Εδώ, δίνεται στο μηχάνημα ένα μεγάλο σύνολο δεδομένων, στο οποίο επισημαίνονται μόνο μερικά σημεία δεδομένων (Rebala et al., 2019). Ο αλγόριθμος θα χρησιμοποιήσει τεχνικές ομαδοποίησης (μη εποπτευόμενη μάθηση) για να προσδιορίσει ομάδες μέσα στο δεδομένο σύνολο δεδομένων και να χρησιμοποιήσει τα λίγα σημεία δεδομένων με ετικέτα σε κάθε ομάδα για να παρέχει ετικέτες σε άλλα σημεία δεδομένων στο ίδιο σύμπλεγμα. Ένα από τα πιο σημαντικά πλεονεκτήματα αυτής της τεχνικής είναι ότι δεν χρειάζεται να ξοδευτεί πολύς χρόνος και προσπάθεια για την επισήμανση κάθε σημείου δεδομένων.

4. **Ενισχυτική μάθηση (Reinforcement Learning)**: Σε αυτήν την κατηγορία ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον (Γεωργούλη, 2015). Ο αλγόριθμος λαμβάνει ανατροφοδότηση με τη μορφή ανταμοιβών ή ποινών με βάση τις ενέργειές του, επιτρέποντάς του να μάθει τις βέλτιστες στρατηγικές. Η προσέγγιση αυτή είναι διαδεδομένη σε τομείς όπως τα παιχνίδια, η ρομποτική και τα αυτόνομα συστήματα. Η ενισχυτική μάθηση είναι ένας από τους πιο συναρπαστικούς τομείς της μηχανικής μάθησης, καθώς είναι χρήσιμη για καταστάσεις που περιλαμβάνουν: (Rebala et al., 2019)

- **Προβλήματα Σχεδιασμού (Planning),** όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.
- **Αλλαγή καταστάσεων**, για παράδειγμα, οδήγηση, παιχνίδι σκάκι, τάβλι κ.λπ. Εδώ, η εξωτερική κατάσταση (ή το παιχνίδι του αντιπάλου) αλλάζει συνεχώς και η απόκριση από το μηχάνημα πρέπει να λαμβάνει υπόψη το αλλαγμένο περιβάλλον.
- **Τεράστιος χώρος κατάστασης**, τα παιχνίδια οδήγησης και πολλών παικτών είναι παραδείγματα. Παιχνίδια όπως το σκάκι έχουν σχεδόν άπειρες πιθανές διαμορφώσεις σανίδων. Δεν είναι δυνατό να παίξετε καλά αυτά τα παιχνίδια χρησιμοποιώντας ωμή βία αναζήτηση για κινήσεις, από μια απαρίθμηση όλων των πιθανών μονοπατιών μέχρι την πρόοδο του παιχνιδιού. Υπάρχουν πάρα πολλά πιθανά μονοπάτια για να τα απαριθμήσουμε.

2.7.3 Ταξινόμηση (Classification)

Η ταξινόμηση είναι ένας από τους κύριους αλγορίθμους στον χώρο της μηχανικής μάθησης. Στην πραγματικότητα, αυτός ο αλγόριθμος αποτελεί τη διαδικασία κατά την οποία τα δεδομένα χωρίζονται σε διακριτές κατηγορίες ή κλάσεις, με βάση τα χαρακτηριστικά τους. Κάθε κατηγορία σχετίζεται με μια ετικέτα, και ο στόχος είναι να εκπαιδεύσουμε ένα μοντέλο που θα μπορεί να προβλέπει τη σωστή κατηγορία για νέα δεδομένα. Το πιθανό σύνολο κατηγοριών είναι επισημασμένο και τα μοντέλα μαθαίνουν γενικά από δεδομένα εκπαίδευσης. Τα μοντέλα ταξινόμησης μπορούν να δημιουργηθούν χρησιμοποιώντας απλά κατώφλια, τεχνικές παλινδρόμησης ή άλλες τεχνικές μηχανικής μάθησης όπως τα Νευρωνικά Δίκτυα, τα Τυχαία Δάση ή τα μοντέλα Markov (Rebala et al., 2019). Η ταξινόμηση είναι ένας

εποπτευόμενος αλγόριθμος εκμάθησης όπου ένα εκπαιδευτικό σύνολο δεδομένων σωστά προσδιορισμένο και επισημασμένο είναι διαθέσιμο.

Τα βασικά βήματα του αλγορίθμου ταξινόμησης περιλαμβάνουν τη συλλογή και προετοιμασία των δεδομένων, την επιλογή ενός κατάλληλου μοντέλου, την εκπαίδευση αυτού του μοντέλου με τα δεδομένα εκπαίδευσης, την αξιολόγηση του μοντέλου με χρήση μετρικών απόδοσης, τη βελτιστοποίηση του μοντέλου εάν απαιτείται, και τέλος τη χρήση του μοντέλου για την πρόβλεψη των κατηγοριών νέων δεδομένων.

Το μοντέλο που δημιουργείται από τα δεδομένα εκπαίδευσης για τον προσδιορισμό της κατηγορίας ή της κλάσης του χαρακτηριστικού ή των δεδομένων εισόδου ονομάζεται ταξινομητής, classifier. Ο ταξινομητής μπορεί να είναι ένας δυαδικός ταξινομητής ή ένας ταξινομητής πολλαπλών κλάσεων. Ένας δυαδικός ταξινομητής προσδιορίζει ότι τα δεδομένα εισόδου ανήκουν σε μία από τις δύο κατηγορίες εξόδου. Για παράδειγμα, η αλληλογραφία που ελήφθη είναι ανεπιθύμητη ή όχι ανεπιθύμητη. Ένας ταξινομητής πολλαπλών κλάσεων προσδιορίζει τα δεδομένα εισόδου σε μία κατηγορία όταν έχει να επιλέξει ανάμεσα σε περισσότερες από δύο κατηγορίες. Για παράδειγμα, η αλληλογραφία που ελήφθη είναι ένα προωθητικό email που αντιπροσωπεύει κάποιο είδος διαφήμισης, προσωπικό email που ελήφθη από φίλους ή συνεργάτες ή ένα ανεπιθύμητο email.

Οι αλγόριθμοι ταξινόμησης χρησιμοποιούνται ευρέως σε πολλούς τομείς, όπως η αναγνώριση προτύπων, η ανάλυση κειμένου, η ανίχνευση απάτης και η ιατρική διάγνωση, μεταξύ άλλων. Η επιλογή του κατάλληλου αλγορίθμου και η καλή προ-επεξεργασία των δεδομένων είναι κρίσιμες για την επίτευξη υψηλής ακρίβειας και απόδοσης στην ταξινόμηση.

2.7.4 Δέντρα Απόφασης (Decision Trees)

Τα Δέντρα Απόφασης είναι μια προσέγγιση εποπτευόμενης μάθησης (supervised learning) που χρησιμοποιείται στη στατιστική, την εξόρυξη δεδομένων και τη μηχανική μάθηση. Σε αυτό το πλαίσιο, ένα δέντρο αποφάσεων ταξινόμησης ή παλινδρόμησης χρησιμοποιείται ως προγνωστικό μοντέλο για την εξαγωγή συμπερασμάτων σχετικά με ένα σύνολο παρατηρήσεων. Σημειώνεται, πως τα δέντρα αποφάσεων συγκαταλέγονται μεταξύ των πιο δημοφιλών αλγορίθμων μηχανικής μάθησης, δεδομένης της ευφυΐας και της απλότητάς τους.

Ένα δέντρο είναι μια άκυκλη κατευθυνόμενη δομή δεδομένων με κόμβους και ακμές που συνδέουν αυτούς τους κόμβους. Ουσιαστικά, είναι ένα δέντρο με κόμβους που αντιπροσωπεύουν ντετερμινιστικές αποφάσεις που βασίζονται σε μεταβλητές και ακμές που αντιπροσωπεύουν τη διαδρομή προς τον επόμενο κόμβο ή έναν κόμβο φύλλου που βασίζεται στην απόφαση. Ο κόμβος φύλλου ή ο τερματικός κόμβος του δέντρου αντιπροσωπεύει μια ετικέτα κλάσης ως έξοδο πρόβλεψης. Επομένως, ο στόχος είναι η δημιουργία ενός μοντέλου που προβλέπει την τιμή μιας επιλεγμένης μεταβλητής με βάση διάφορες μεταβλητές εισόδου.

Για παράδειγμα, δεδομένου ενός αντικειμένου, μας ζητείται να ταξινομήσουμε το αντικείμενο ως μήλο ή πορτοκάλι ή κανένα από τα δύο, απλώς κάνοντας ερωτήσεις σχετικά με το αντικείμενο. Οι απαντήσεις σε μια σειρά ερωτήσεων σχετικά με το αντικείμενο θα μας οδηγήσουν στο συμπέρασμα της ταξινόμησης του αντικειμένου. Ορισμένα δείγματα ερωτήσεων που μπορούμε να κάνουμε είναι τα εξής: είναι το αντικείμενο βρώσιμο, είναι το αντικείμενο στρογγυλό, είναι το χρώμα του αντικειμένου πορτοκαλί, μπορούμε να ξεφλουδίσουμε το αντικείμενο με το χέρι, κ.λπ. Κάθε κόμβος αντιπροσωπεύει μια ερώτηση και η άκρη από τον κόμβο που οδηγεί στον επόμενο κόμβο αντιπροσωπεύει τη διαδρομή που ακολουθήθηκε με βάση την απάντηση. Εάν οι ερωτήσεις απαντηθούν σωστά, τότε ο τελευταίος κόμβος ή τερματικός κόμβος θα ολοκληρώσει την κλάση του αντικειμένου ως πορτοκάλι ή μήλο ή κανένα από τα δύο (Rebala et al., 2019).

Ένα Δέντρο Αποφάσεων διαμορφώνεται σε μια απλή σειρά ερωτήσεων που οδηγούν σειριακά σε μια απάντηση που ταιριάζει καλύτερα στα δεδομένα που χρησιμοποιούνται στην εκπαίδευση. Τα χειροποίητα δέντρα αποφάσεων χρησιμοποιήθηκαν συνήθως στη μηχανική λειτουργιών για τον προσδιορισμό της σημασίας της μεταβλητής στην απόφαση ή για την πρόβλεψη του αποτελέσματος. Οι ερωτήσεις που τίθενται σε κάθε σημείο απόφασης ή κόμβο ενός δέντρου οδηγούν σε ένα μονοπάτι που χρησιμοποιεί μοντέλα "if a then x else y". Η κατασκευή Δέντρου Αποφάσεων είναι διαισθητική και μπορεί εύκολα να κατασκευαστεί για μικρό αριθμό στοιχείων απόφασης με το χέρι. Για μεγάλο όγκο δεδομένων, το Δέντρο Αποφάσεων μπορεί να κατασκευαστεί χρησιμοποιώντας την τεχνική bagging (δειγματοληψία ενός υποσυνόλου δεδομένων εκπαίδευσης με αντικατάσταση) με τους κανόνες που εξάγονται από τα δεδομένα (Kubat, 2017).

Ένα δέντρο απόφασης δημιουργείται ακολουθώντας τρία βήματα:

1) Δημιουργούμε τη ρίζα με μεταβλητές που είναι πιο σημαντικές.
2) Δημιουργούμε μια απόφαση με βάση την υψηλότερη κατανομή πληροφοριών.
3) Κατασκευάζουμε αναδρομικά τους κόμβους και αποφασίζουμε χρησιμοποιώντας το βήμα ένα και το βήμα δύο έως ότου δεν μπορεί να διαχωριστεί καμία πληροφορία στον κόμβο άκρης.

Όταν εμπλέκεται μεγάλος αριθμός μεταβλητών, η κύρια πρόκληση για τη δημιουργία ενός Δέντρου Αποφάσεων είναι η εύρεση μεταβλητής ή συνδυασμού μεταβλητών μεγαλύτερης σημασίας σε κάθε έναν από τους κόμβους. Αυτή η επιλογή μεταβλητής, που ονομάζεται επίσης επιλογή χαρακτηριστικών, εκτελείται γενικά χρησιμοποιώντας το κέρδος πληροφοριών ή το κριτήριο ακαθαρσίας Gini. Το κέρδος πληροφοριών χρησιμοποιείται όταν οι μεταβλητές είναι κατηγορικές, δηλαδή όταν οι τιμές των μεταβλητών εμπίπτουν σε κλάσεις ή κατηγορίες και δεν έχουν λογική σειρά, για παράδειγμα, τύποι φρούτων. Η ακαθαρσία Gini χρησιμοποιείται όταν οι μεταβλητές τιμές είναι συνεχείς, δηλαδή οι τιμές είναι αριθμητικές, για παράδειγμα, η ηλικία ενός ατόμου.

Κέρδος πληροφοριών

Χρησιμοποιώντας τη θεωρία πληροφοριών, υπολογίζουμε την ποσότητα των πληροφοριών που περιέχονται σε κάθε μεταβλητή. Ένα βασικό μέτρο στη θεωρία της πληροφορίας είναι η

εντροπία. Η εντροπία ποσοτικοποιεί το ποσό της αβεβαιότητας που εμπλέκεται στην τιμή της τυχαίας μεταβλητής. Για παράδειγμα, σε δυαδικά αποτελέσματα 0 ή 1, εάν όλα τα αποτελέσματα είναι 1, τότε η εντροπία είναι μηδέν καθώς δεν υπάρχει αβεβαιότητα στην πρόβλεψη του αποτελέσματος. Από την άλλη πλευρά, εάν το αποτέλεσμα του 0 ή του 1 είναι ίσο, τότε η αβεβαιότητα ή η εντροπία είναι υψηλή (Rebala et al., 2019).

Κριτήριο ακαθαρσίας Gini

Το κριτήριο ακαθαρσίας Gini είναι ένα μέτρο του πόσο συχνά ένα τυχαία επιλεγμένο στοιχείο από το σύνολο επισημαίνεται λανθασμένα εάν είχε επισημανθεί σύμφωνα με την κατανομή των ετικετών στο υποσύνολο. Η υψηλότερη ακαθαρσία Gini αναφέρεται σε μεγαλύτερη πιθανότητα εσφαλμένης ταξινόμησης αντίστροφα, και η χαμηλότερη ακαθαρσία αναφέρεται σε μικρότερη πιθανότητα εσφαλμένης ταξινόμησης. Κατά την κατασκευή ενός διαχωρισμού Δέντρου Αποφάσεων, ο στόχος είναι να διαχωριστεί με τη χαμηλότερη σταθμισμένη τιμή ακαθαρσίας Gini για τα θυγατρικά δέντρα. Αυτό μεταφράζεται σε διαχωρισμό με τη μεγαλύτερη μείωση της σταθμισμένης ακαθαρσίας Gini για θυγατρικούς κόμβους από την ακαθαρσία Gini του γονικού κόμβου. Με άλλα λόγια, η διάσπαση θα έχει ως αποτέλεσμα τη μεγαλύτερη δυνατή ομοιογένεια στον κόμβο του παιδιού (Kubat, 2017).

2.7.5 Τυχαία Δάση (Random Forest)

Τα τυχαία δάση είναι συλλογή δέντρων απόφασης, δηλαδή είναι ένας ταξινομητής συνόλου που χρησιμοποιεί πολλά μοντέλα δέντρων αποφάσεων και είναι ένας τρόπος υπολογισμού του μέσου όρου πολλαπλών δέντρων αποφάσεων που έχουν δημιουργηθεί από διαφορετικά μέρη του σετ εκπαίδευσης με στόχο τη μείωση της υπερβολικής προσαρμογής από ένα μόνο δέντρο απόφασης. Η αυξημένη απόδοση έρχεται με κόστος κάποιας απώλειας ερμήνευσης και ακριβούς προσαρμογής στα δεδομένα προπόνησης (Kubat, 2017). Τα τυχαία δάση ή τυχαία δάση απόφασης (random decision forest) είναι μια μέθοδος μάθησης συνόλου για ταξινόμηση, παλινδρόμηση αλλά και για άλλες εργασίες, που λειτουργεί με την κατασκευή ενός πλήθους δέντρων απόφασης κατά τη διάρκεια της εκπαίδευσης. Πιο συγκεκριμένα, για προβλήματα ταξινόμησης, η έξοδος του τυχαίου δάσους είναι η κλάση που επιλέγεται από τα περισσότερα δέντρα, ενώ αντίστοιχα για εργασίες παλινδρόμησης, επιστρέφεται ο μέσος όρος ή η μέση πρόβλεψη των μεμονωμένων δέντρων (Ho, Tin Kam., 1995).

Τα τυχαία δάση χρησιμοποιούν το bagging για να κατασκευάσουν πολλαπλά δέντρα απόφασης. Το Bagging ή bootstrap agregation είναι μια τεχνική δειγματοληψίας ενός υποσυνόλου δεδομένων εκπαίδευσης με αντικατάσταση και κατασκευάζει το μοντέλο με βάση το δειγματοληπτικό σύνολο εκπαίδευσης. Για ένα μεγάλο σύνολο δεδομένων D, αναμένεται να έχει ~ 62,3% μοναδικών δειγμάτων από το D σε καθένα από τα υποσύνολα.

Τα τυχαία δάση αναπτύσσουν πολλά δέντρα απόφασης λαμβάνοντας δειγματοληψία από το σύνολο δεδομένων εισόδου D για ένα σύνολο χαρακτηριστικών αντί για τα δείγματα δεδομένων εκπαίδευσης. Η διαδικασία ονομάζεται επίσης χαρακτηριστική συσκευασία. Εάν ένα από τα χαρακτηριστικά έχει υψηλή συσχέτιση, τότε αυτό το χαρακτηριστικό θα επιλεγεί

από πολλά διαφορετικά δέντρα. Αυτό προκαλεί τη συσχέτιση των δέντρων και έτσι αυξάνει την προκατάληψη του αλγορίθμου, μειώνοντας την ακρίβεια (Rebala et al., 2019).

2.7.6 Support Vector Machines (SVM)

Τα SVM εισήχθησαν στις αρχές της δεκαετίας του 1990 και ήταν επιτυχείς στην εφαρμογή σε προβλήματα ταξινόμησης και παλινδρόμησης πραγματικού κόσμου. Ένα από τα πλεονεκτήματα των SVM είναι ότι μπορούν να λειτουργήσουν με αραιά δεδομένα και τα μοντέλα δημιουργούνται με σχετικά μικρό σύνολο δειγμάτων. Τα SVM παρέχουν έναν συμβιβασμό μεταξύ παραμετρικών και μη παραμετρικών προσεγγίσεων. Οι παραμετρικές προσεγγίσεις χρησιμοποιούν παραμέτρους για τη μοντελοποίηση, παρόμοιες με τη γραμμική παλινδρόμηση, ενώ τα μη παραμετρικά μοντέλα όπως τα δέντρα αποφάσεων περιλαμβάνουν άμεσα δεδομένα εκπαίδευσης και δεν εξαρτώνται από παραμέτρους. Τα SVM είναι δυαδικοί γραμμικοί ταξινομητές, δηλ. δημιουργούν ένα όριο υπερεπιπέδου για να διαχωρίσουν τα σημεία δεδομένων σε δύο κατηγορίες. Μπορούν να χειριστούν μόνο σημεία δεδομένων για τα οποία υπάρχει όριο υπερεπιπέδου. Τέτοια σημεία δεδομένων ονομάζονται γραμμικά διαχωρίσιμα. Τα SVM μπορούν να χειριστούν σημεία δεδομένων που δεν διαχωρίζονται γραμμικά αντιστοιχίζοντας σημεία δεδομένων σε χώρο υψηλότερων διαστάσεων χρησιμοποιώντας συναρτήσεις με ειδικές ιδιότητες που ονομάζονται πυρήνες (Kubat, 2017).

Ο ταξινομητής SVM δημιουργεί ένα υπερεπίπεδο διαστάσεων N - 1 για n διανύσματα χαρακτηριστικών διαστάσεων (για N = 2, το υπερεπίπεδο είναι μια γραμμή) για να διαχωρίσει τα δεδομένα σε δύο κλάσεις. Η γραμμή ταξινομητή μπορεί να αναπαρασταθεί από την εξίσωση:

$$y = w * f(x) + b,$$

όπου το f(x) είναι το διάνυσμα χαρακτηριστικών, w είναι το βάρος που αποδίδεται στο διάνυσμα χαρακτηριστικών και b είναι ο όρος μεροληψίας (Rebala et al., 2019).

Οι πυρήνες μπορούν να επιλεγούν με βάση το σύνολο δεδομένων και τα χαρακτηριστικά του συνόλου δεδομένων. Συνήθως, δεν είναι προφανές ποιος πυρήνας λειτουργεί καλύτερα. Είναι συνήθως ωφέλιμο να ξεκινήσουμε με απλούς πυρήνες και να επεξεργαστούμε μέχρι πολύπλοκους πυρήνες. Με βάση αυτή την προσέγγιση, θα ξεκινούσαμε με έναν γραμμικό πυρήνα και εάν η ταξινόμηση δεν αποδίδει καλά, τότε επιλέγουμε έναν μη γραμμικό πυρήνα.

Ο πυρήνας ακτινικής προκατάληψης (RBF) είναι ένας από τους συνήθως χρησιμοποιούμενους πυρήνες στο SVM. Οι πολυωνυμικές και σιγμοειδείς συναρτήσεις πυρήνα είναι μερικοί από τους άλλους συνήθως χρησιμοποιούμενους πυρήνες.

Η διασταυρούμενη επικύρωση είναι ένας καλός τρόπος για να προσδιορίσουμε ποιος από τους πυρήνες λειτουργεί καλύτερα για τα δεδομένα. Το SVM είναι ιδιαίτερα χρήσιμο όταν το σύνολο δεδομένων εισόδου είναι μικρό και ο αριθμός των χαρακτηριστικών είναι σχετικά υψηλός. Το SVM μπορεί να μάθει με μικρό όγκο δεδομένων για τη δημιουργία ενός ορίου απόφασης (Rebala et al., 2019).

2.7.7 Νευρωνικό δίκτυο (Neural Network)

Στη μηχανική μάθηση, ένα νευρωνικό δίκτυο (επίσης τεχνητό νευρωνικό δίκτυο ή νευρωνικό δίκτυο, συντομογραφία ANN ή NN) είναι ένα μοντέλο εμπνευσμένο από τη νευρωνική οργάνωση που βρίσκεται στα βιολογικά νευρωνικά δίκτυα στους εγκεφάλους των ζώων [Hardesty, Larry, 2017). Ένα νευρωνικό δίκτυο αποτελείται από συνδεδεμένες μονάδες ή κόμβους που ονομάζονται τεχνητοί νευρώνες, οι οποίοι μοντελοποιούν χαλαρά τους νευρώνες ενός εγκεφάλου. Αυτοί συνδέονται με ακμές, οι οποίες μοντελοποιούν τις συνάψεις σε έναν εγκέφαλο. Ένας τεχνητός νευρώνας λαμβάνει σήματα από συνδεδεμένους νευρώνες, στη συνέχεια τα επεξεργάζεται και στέλνει ένα σήμα σε άλλους συνδεδεμένους νευρώνες. Το "σήμα" είναι ένας πραγματικός αριθμός και η έξοδος κάθε νευρώνα υπολογίζεται από κάποια μη γραμμική συνάρτηση του αθροίσματος των εισόδων του, που ονομάζεται συνάρτηση ενεργοποίησης. Οι νευρώνες και οι ακμές έχουν συνήθως ένα βάρος που προσαρμόζεται καθώς προχωρά η μάθηση. Το βάρος αυξάνει ή μειώνει την ισχύ του σήματος σε μια σύνδεση.

Συνήθως, οι νευρώνες συγκεντρώνονται σε επίπεδα. Διαφορετικά στρώματα μπορούν να εκτελούν διαφορετικούς μετασχηματισμούς στις εισόδους τους. Τα σήματα ταξιδεύουν από το πρώτο στρώμα (το στρώμα εισόδου) στο τελευταίο στρώμα (το στρώμα εξόδου), περνώντας ενδεχομένως από πολλαπλά ενδιάμεσα στρώματα (κρυφά στρώματα). Ένα δίκτυο ονομάζεται συνήθως βαθύ νευρωνικό δίκτυο αν έχει τουλάχιστον 2 κρυφά στρώματα (Bishop, Christopher M., and Nasser M. Nasrabad, 2006).

Τα τεχνητά νευρωνικά δίκτυα χρησιμοποιούνται για προγνωστική μοντελοποίηση, προσαρμοστικό έλεγχο και άλλες εφαρμογές όπου μπορούν να εκπαιδευτούν μέσω ενός συνόλου δεδομένων. Χρησιμοποιούνται επίσης για την επίλυση προβλημάτων στην τεχνητή νοημοσύνη. Τα δίκτυα μπορούν να μαθαίνουν από την εμπειρία και μπορούν να εξάγουν συμπεράσματα από ένα πολύπλοκο και φαινομενικά άσχετο σύνολο πληροφοριών.

Τα νευρωνικά δίκτυα συνήθως εκπαιδεύονται μέσω εμπειρικής ελαχιστοποίησης του κινδύνου. Αυτή η μέθοδος βασίζεται στην ιδέα της βελτιστοποίησης των παραμέτρων του δικτύου για την ελαχιστοποίηση της διαφοράς, ή εμπειρικού κινδύνου, μεταξύ της προβλεπόμενης εξόδου και των πραγματικών τιμών-στόχων σε ένα δεδομένο σύνολο δεδομένων (Vapnik, 1998). Μέθοδοι που βασίζονται στην κλίση, όπως η οπισθοδιάδοση, χρησιμοποιούνται συνήθως για την εκτίμηση των παραμέτρων του δικτύου (Vapnik, 1998). Κατά τη φάση της εκπαίδευσης, τα ANN μαθαίνουν από τα επισημασμένα δεδομένα εκπαίδευσης ενημερώνοντας επαναληπτικά τις παραμέτρους τους για την ελαχιστοποίηση μιας καθορισμένης συνάρτησης απώλειας.[Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, 2016.] Αυτή η μέθοδος επιτρέπει στο δίκτυο να γενικεύει σε αόρατα δεδομένα.

3. Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας περιλαμβάνει τα παρακάτω βήματα:

1. Μελέτη Πεδίων: Αρχικά, ως πρώτο βήμα της εργασίας ήταν η μελέτη των γνωστικών πεδίων. Μελετήθηκαν σε βάθος έννοιες όπως η επιστήμη των δεδομένων, η τεχνητή νοημοσύνη και η μηχανική μάθηση. Επίσης, μελετήθηκαν τα πεδία εφαρμογής της τεχνητής νοημοσύνης καθώς και ένα εργαλείο που μπορούμε να χρησιμοποιήσουμε, το Microsoft Azure. Ακόμη, είδαμε αναλυτικά την χρήση αυτής της πλατφόρμας σε προβλήματα που μπορεί να αντιμετωπίσει μια επιχείρηση. Αναφορικά με την μηχανική μάθηση μελετήθηκαν τα μοντέλα μάθησης και έπειτα είδαμε μερικούς αλγορίθμους που μπορούμε να χρησιμοποιήσουμε.

2. Μελέτη πλατφόρμας Microsoft Azure: Έπειτα, ερευνήθηκε η πλατφόρμα Microsoft Azure, οι οποία καλύπτει τις ανάγκες της παρούσας διπλωματικής. Για το σκοπό αυτό μελετήθηκε ο τρόπος με τον οποίο λειτουργεί και πως μπορούμε να αξιοποιήσουμε τις δυνατότητες που προσφέρει. Σε θεωρητικό και πρακτικό επίπεδο είδαμε το πεδίο χρήσης του εργαλείου καθώς και πως μπορεί να φανεί χρήσιμο στην καθημερινότητα μιας επιχείρησης ή και ενός οργανισμού.

3. Σχεδιασμός και Υλοποίηση Σεναρίων: Τρίτο βήμα, αφού ολοκληρώθηκε σε ικανοποιητικό βαθμό η μελέτη του εργαλείου, ήταν ο σχεδιασμός και η υλοποίηση των σεναρίων. Δημιουργήθηκαν διαφορετικά σενάρια, τα οποία βασίστηκαν κυρίως στις δυνατότητες που προσέφερε η πλατφόρμα του Azure. Στο πρώτο σενάριο δημιουργήθηκε μια αναζήτηση και ανάλυση δεδομένων στο οποίο βασικός άξονα ήταν η γνωριμία του χρήστη με τις τεχνολογίες του Data Lake και του Serverless SQL Pool. Σημειώνεται πως απευθύνεται σε μη έμπειρους ή εξοικειωμένους χρήστες με το Azure ή αντίστοιχα με άλλες πλατφόρμες και για αυτό το λόγο είναι σε μορφή οδηγιών. Στο δεύτερο σενάριο που υλοποιήθηκε είδαμε πως μπορούμε να αναπτύξουμε και να εκπαιδεύσουμε ένα μοντέλο παλινδρόμησης για να εξάγουμε αξιόπιστες προβλέψεις για την τιμή που ερευνούμε. Τέλος, στο τρίτο σενάριο δημιουργήθηκε ένα μοντέλο ταξινόμησης κυρίως με την χρήση ενός αλγορίθμου ταξινόμησης. Στην συνέχεια εκτελέστηκαν κι άλλοι αλγόριθμοι ταξινόμησης με σκοπό να αξιολογήσουμε καλύτερα τα αποτελέσματα.

4. Αξιολόγηση πλατφόρμας και συμπεράσματα: Μετά την υλοποίηση των σεναρίων και την τριβή με την πλατφόρμα του Azure, επόμενο βήμα είναι η αξιολόγηση του, ως προς την χρηστικότητα του, τις δυνατότητες που παρέχει, την ευκολία χρήσης του, καθώς και την πολυπλοκότητα των δράσεων που μπορούν να λάβουν χώρα. Επιπλέον της αξιολόγησης, στο τελευταίο κεφάλαιο προχωρήσαμε στα συμπεράσματα της εργασίας, τόσο ως προς την πλατφόρμα όσο και ως προς την γενικότερη χρησιμότητα του χώρου της τεχνητής νοημοσύνης και της μηχανικής μάθησης.

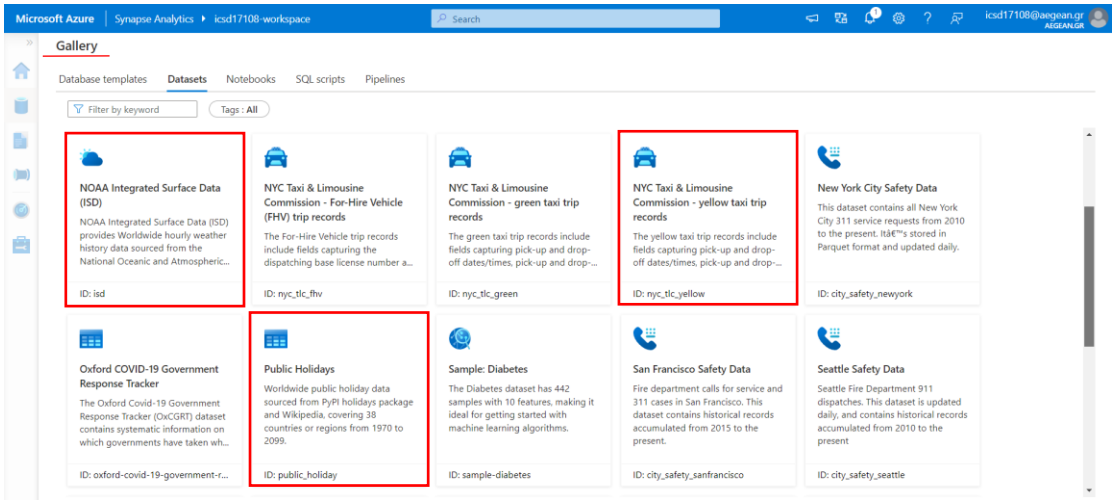
4. Υλοποίηση Σεναρίων

Στο κεφάλαιο αυτό θα υλοποιήσουμε μερικά σενάρια. Βασικό εργαλείο για την δημιουργία των σεναρίων αυτών θα είναι το Microsoft Azure. Το Microsoft Azure είναι η cloud πλατφόρμα της Microsoft με την οποία μπορούμε να αντικαταστήσουμε τον κλασικό server, δίνοντας την δυνατότητα παροχής ευέλικτων υπηρεσιών καθώς και την εξυπηρέτηση χρηστών που θέλουν να συνεχίσουν να εργάζονται από οπουδήποτε και οποτεδήποτε, απρόσκοπτα και με ασφάλεια. Κάθε σενάριο θα έχει ένα διαφορετικό βαθμό δυσκολίας στην υλοποίηση του ώστε να μπορέσουμε να δούμε αρκετές διαφορετικές πλευρές από την χρήση του εργαλείου αυτού. Τα δεδομένα που θα χρησιμοποιηθούν είναι κυρίως δεδομένα που παρέχονται από την επίσημη βάση δεδομένων του εργαλείου για εξάσκηση με αυτά.

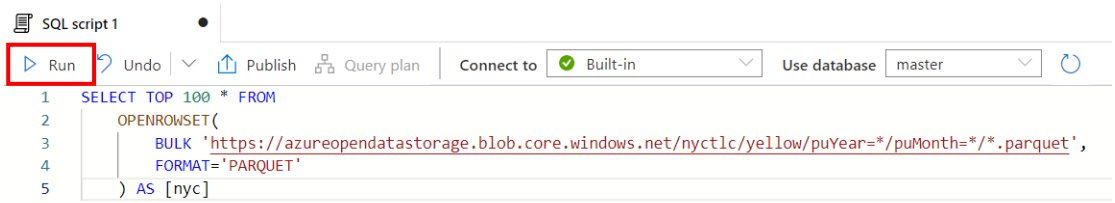
4.1 Σενάριο 1^{ο} | Αναζήτηση και ανάλυση δεδομένων σε Data Lakes με Serverless SQL Pool

Σε αυτό το σενάριο, θα δούμε πώς εκτελείται μια διερευνητική ανάλυση δεδομένων χρησιμοποιώντας Serverless SQL Pool. Ο σκοπός μας είναι να συνδυάσουμε διαφορετικά σύνολα δεδομένων για την άμεση και εύκολη εξαγωγή πληροφοριών ή συμπερασμάτων. Στην συγκεκριμένη περίπτωση θα χρησιμοποιήσουμε τα έτοιμα σύνολα δεδομένων (Open Datasets) που υπάρχουν στην πλατφόρμα Azure και στην συνέχεια θα οπτικοποιήσουμε τα παραγόμενα αποτελέσματα στο Synapse Studio. Αξίζει να σημειωθεί πως από την στιγμή που τα δεδομένα αποθηκεύονται σε μορφή αρχείου Parquet, τότε η αυτόματη εξαγωγή αποτελεσμάτων πραγματοποιείται από την πλατφόρμα. Ακόμη, μπορούμε να αναζητήσουμε οποιαδήποτε δεδομένα χωρίς να αναφέρουμε ονομαστικά τους τύπους δεδομένων όλων των στηλών που υπάρχουν στα διαφορετικά αρχεία. Επίσης, μπορούμε να χρησιμοποιήσουμε τον μηχανισμό εικονικών στηλών και τη λειτουργία διαδρομής αρχείου για να φιλτράρουμε ένα συγκεκριμένο υποσύνολο αρχείων που μας ενδιαφέρει.

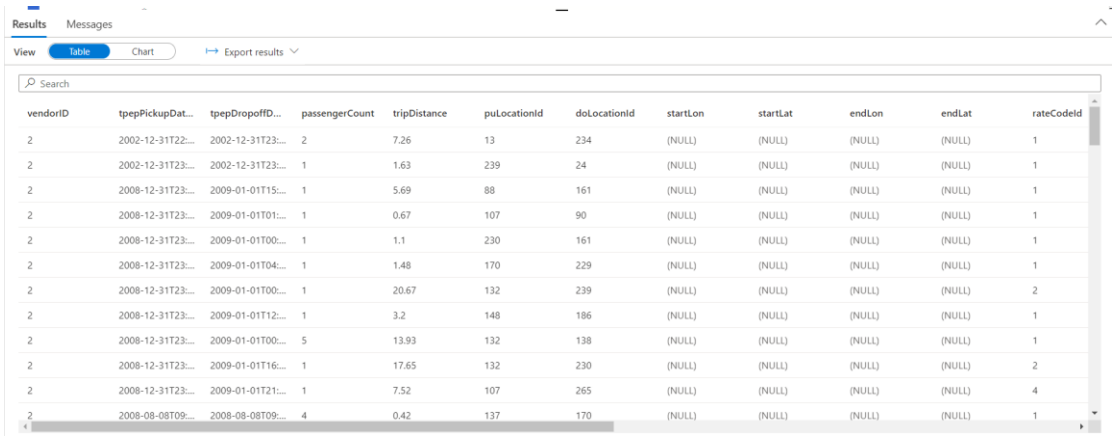
Το βασικό Dataset που θα χρησιμοποιήσουμε αφορά ένα σύνολο δεδομένων σχετικά με τα ταξί της Νέας Υόρκης (NYC), το οποίο περιέχει ημερομηνίες/ώρες παραλαβής και αποβίβασης, τοποθεσίες παραλαβής και αποβίβασης, αποστάσεις διαδρομής, αναλυτικές τιμές ναύλων, τύποι τιμών, τύποι πληρωμής και καταμέτρηση επιβατών σύμφωνα με τον οδηγό. Έπειτα θα συνδυάσουμε το παραπάνω dataset με το σύνολο δεδομένων των επίσημων αργιών και των καιρικών δεδομένων. Όλα τα παραπάνω open datasets μπορούμε να τα βρούμε στην βιβλιοθήκη (Gallery) της πλατφόρμας.



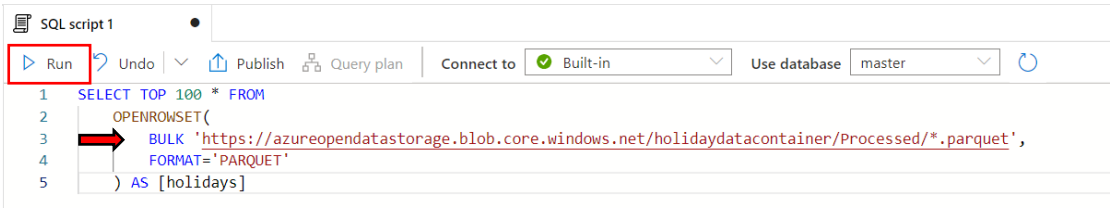
Σε αυτό το σενάριο θα χρησιμοποιήσουμε κώδικα Python για τα queries μας. Αρχικά για να εξοικειωθούμε με τα δεδομένα από τα ταξί της Νέας Υόρκης, εκτελούμε το ακόλουθο ερώτημα, με το οποίο αναζητούμε τις πρώτες εκατό εγγραφές του αρχείου.



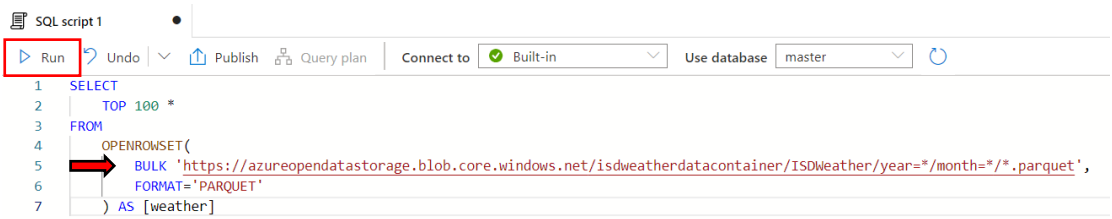
Εφόσον ολοκληρωθεί επιτυχώς το αίτημα μας τότε θα εμφανιστούν τα αποτελέσματα της αναζήτησης, δηλαδή οι πρώτες 100 εγγραφές του αρχείου με όλα τα δεδομένα από τα ταξί της Νέας Υόρκης, όπως φαίνονται στον παρακάτω πίνακα.



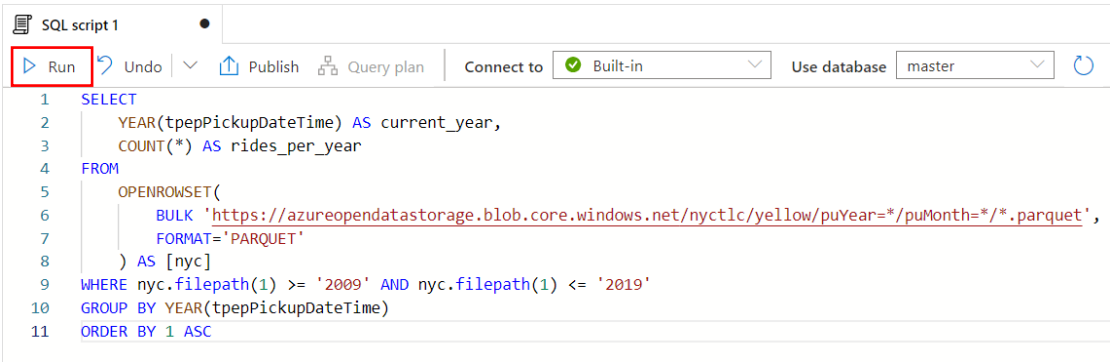
Σε περίπτωση που θέλουμε να μάθουμε περισσότερες πληροφορίες για τη σημασία των επιμέρους στηλών τότε μπορούμε να πάμε στις περιγραφές των συνόλων δεδομένων όπου γράφονται όλες αναλυτικά. Με παρόμοιο τρόπο θα κάνουμε αναζήτηση στο σύνολο δεδομένων για τις επίσημες αργίες χρησιμοποιώντας το παρακάτω τροποποιημένο ερώτημα. Αυτή την φορά αναζητούμε τις πρώτες εκατό εγγραφές του αρχείου με τις επίσημες αργίες. Το μόνο που αλλάζει στον κώδικα (γραμμή 3) σε σχέση με το προηγούμενο ερώτημα είναι η διεύθυνση του αρχείου.



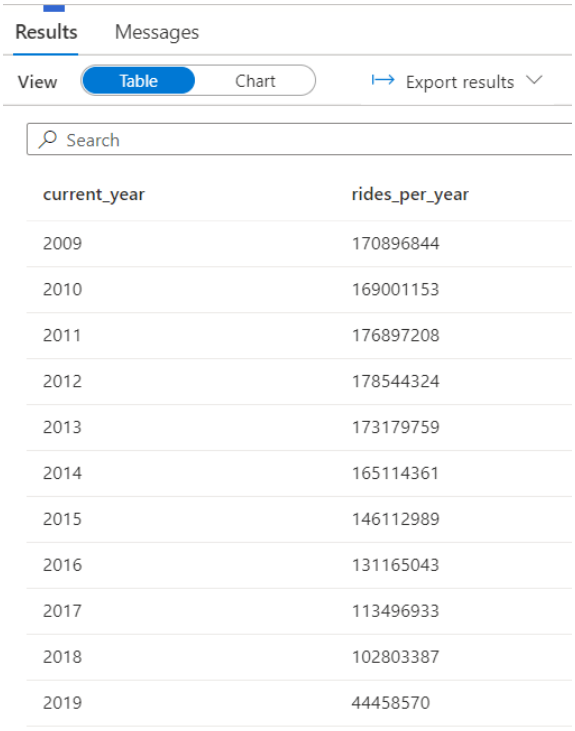
Αντίστοιχα, εκτελούμε την ίδια αναζήτηση και στο σύνολο δεδομένων για τα καιρικά δεδομένα χρησιμοποιώντας το παρακάτω τροποποιημένο ερώτημα.



Στην συνέχεια θα κάνουμε ανάλυση με βάση τις χρονοσειρές, την εποχικότητα και τις ακραίες τιμές. Επομένως, θα συνοψίσουμε τον ετήσιο αριθμό των διαδρομών με ταξί για μια δεκαετία, από το 2009 έως το 2019, χρησιμοποιώντας το ακόλουθο ερώτημα.

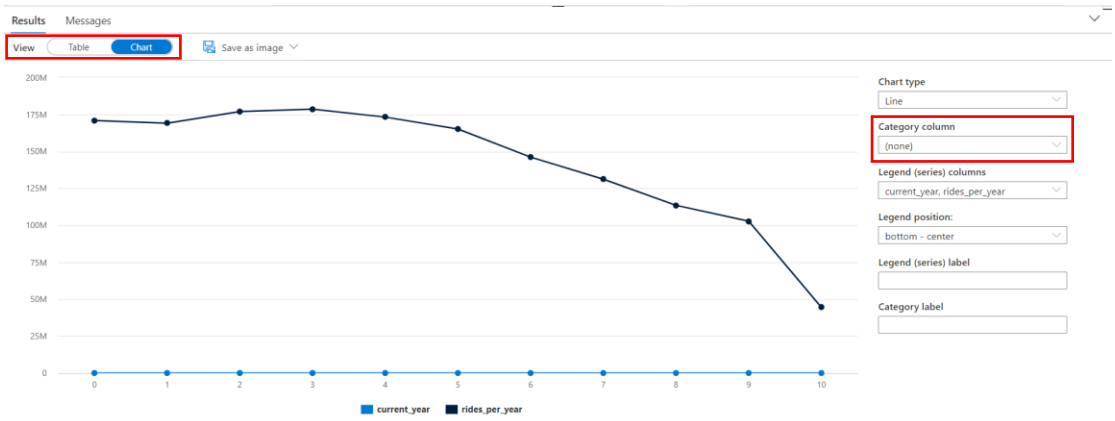


Το ακόλουθο απόσπασμα δείχνει το αποτέλεσμα για τον ετήσιο αριθμό διαδρομών με ταξί, δηλαδή πόσα συνολικά δρομολόγια εκτέλεσαν τα ταξί της Νέας Υόρκης κάθε χρονιά για το διάστημα που ορίσαμε, στην περίπτωση μας μια δεκαετία. Ο πίνακας αποτελεσμάτων περιέχει δύο στήλες, την «current_year» που υποδηλώνει το ημερολογιακό έτος και την «rides_per_year» που υποδηλώνει της διαδρομές ανά έτος.

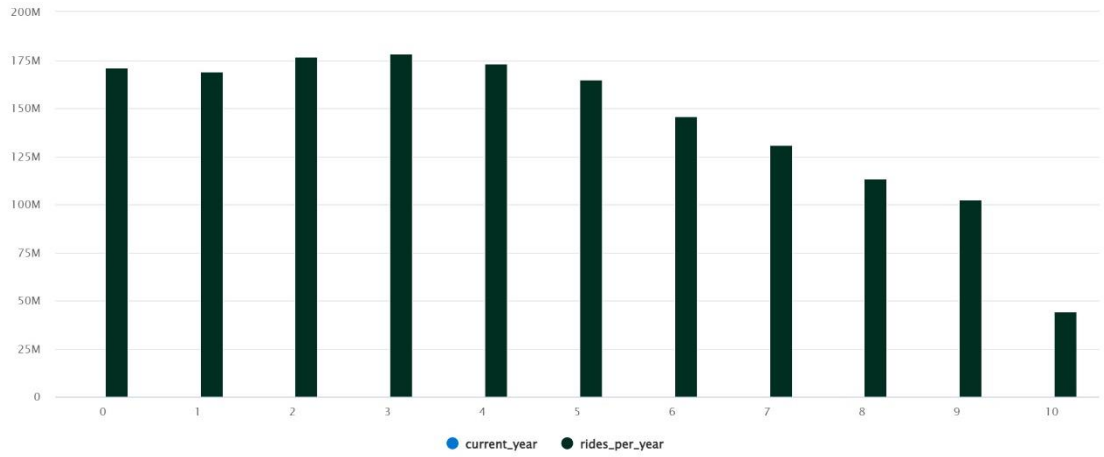


current_year	rides_per_year
2009	170896844
2010	169001153
2011	176897208
2012	178544324
2013	173179759
2014	165114361
2015	146112989
2016	131165043
2017	113496933
2018	102803387
2019	44458570

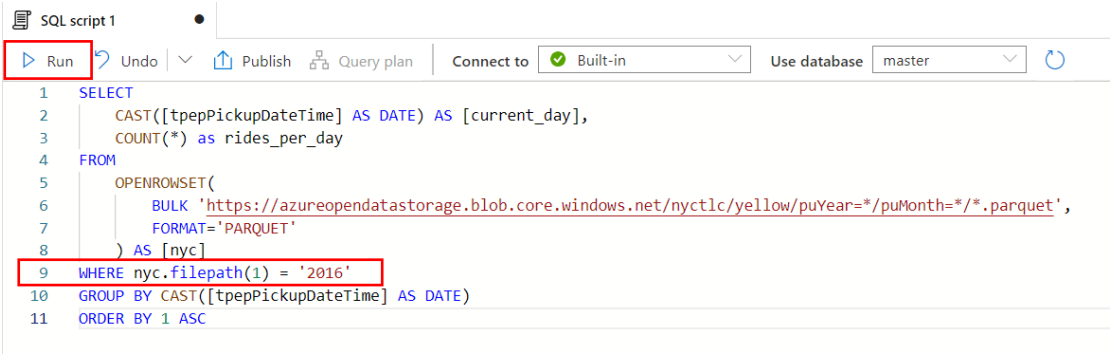
Εναλλακτικά, τα παραπάνω δεδομένα μπορούν να απεικονιστούν στο Synapse Studio μεταβαίνοντας από την προβολή πίνακα στην προβολή διαγράμματος. Υπάρχει η δυνατότητα να επιλέξουμε μεταξύ διαφορετικών τύπων διαγραμμάτων, όπως Area, Bar, Column, Line, Pie και Scatter.



Σε αυτή την περίπτωση, σχεδιάζουμε το διάγραμμα στήλης με την κατηγορία στήλης «Category» να έχει οριστεί ως «current_year», όπως φαίνεται στην παρακάτω εικόνα.



Από αυτή την απεικόνιση, μπορούμε να δούμε ξεκάθαρα μια τάση μείωσης του αριθμού των διαδρομών με την πάροδο των ετών. Κατά πάσα πιθανότητα, η μείωση αυτή οφείλεται στην πρόσφατη αυξημένη δημοτικότητα των εταιρειών κοινής χρήσης διαδρομών. Έπειτα, μπορούμε να εστιάσουμε την ανάλυση σε ένα μόνο έτος για να εξάγουμε περισσότερες πληροφορίες. Ας ορίσουμε για παράδειγμα το 2016, όπου με το ακόλουθο ερώτημα αναζητούμε τον ημερήσιο αριθμό των διαδρομών κατά τη διάρκεια αυτού του έτους.



Το ακόλουθο απόσπασμα δείχνει το αποτέλεσμα σε πίνακα για αυτό το ερώτημα.

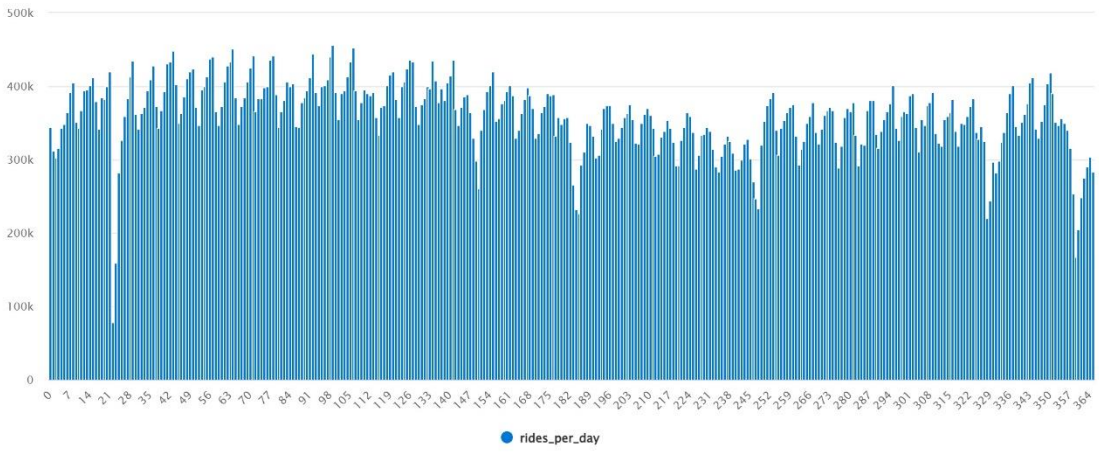
Results Messages

View Table Chart Export results

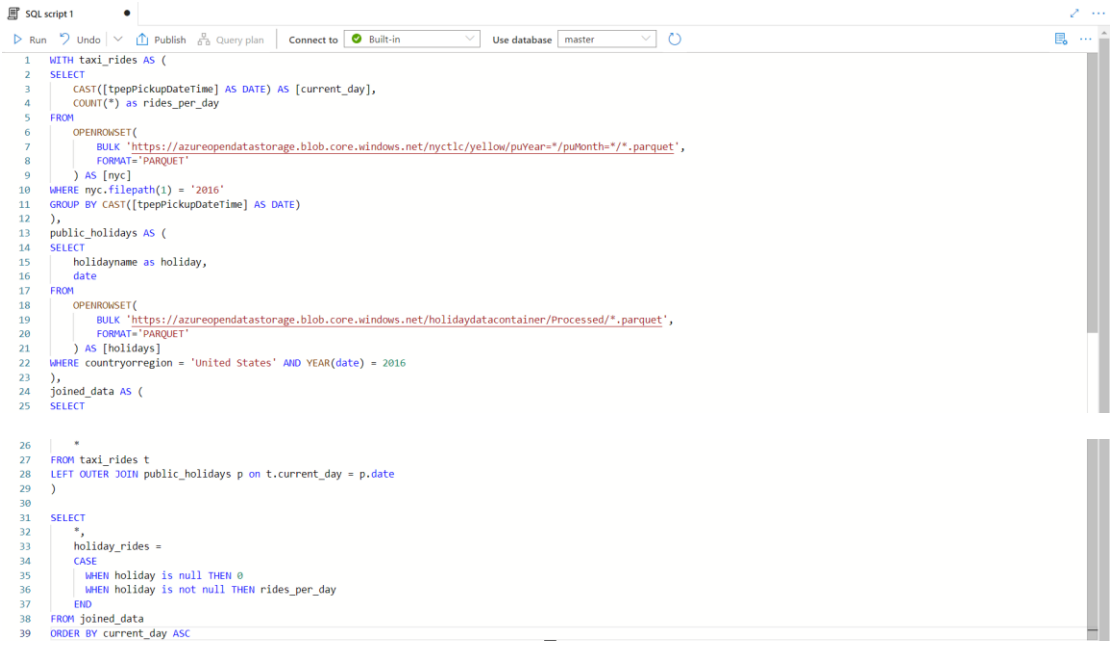
Search

current_day	rides_per_day
2016-01-01T00:00:00.0000000	345037
2016-01-02T00:00:00.0000000	312831
2016-01-03T00:00:00.0000000	302878
2016-01-04T00:00:00.0000000	316171
2016-01-05T00:00:00.0000000	343251
2016-01-06T00:00:00.0000000	348516
2016-01-07T00:00:00.0000000	364894
2016-01-08T00:00:00.0000000	392070
2016-01-09T00:00:00.0000000	405825
2016-01-10T00:00:00.0000000	351788
2016-01-11T00:00:00.0000000	342651
2016-01-12T00:00:00.0000000	367390

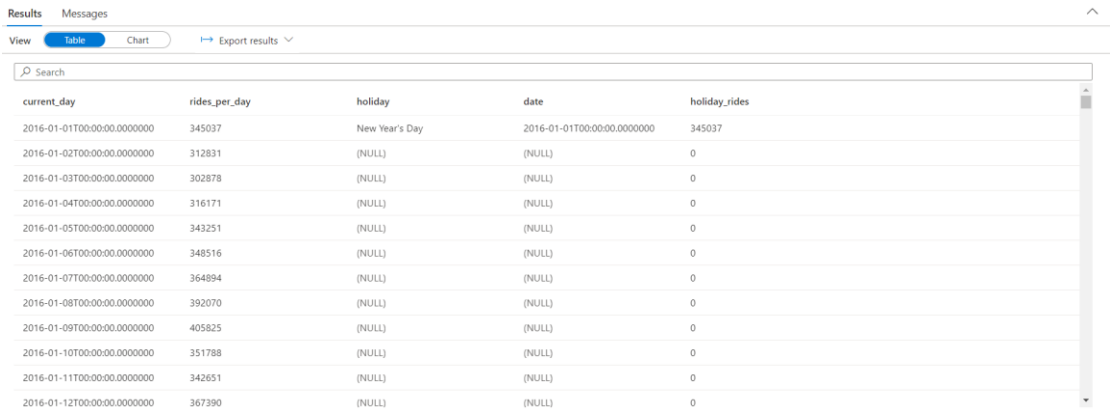
Αντίστοιχα, μπορούμε να απεικονίσουμε τα δεδομένα σχεδιάζοντας το διάγραμμα στήλης με τη κατηγορία στήλης (άξονας Y) να έχει οριστεί ως «current_day» και τη στήλη Υπόμνημα (άξονας X) να έχει οριστεί ως «rides_per_day».



Από το διάγραμμα, μπορούμε να δούμε ότι υπάρχει ένα εβδομαδιαίο μοτίβο, με τα Σάββατα ως ημέρα αιχμής. Σημειώνεται πως κατά τους καλοκαιρινούς μήνες, υπάρχουν λιγότερες διαδρομές με ταξί λόγω των διακοπών. Επίσης, παρατηρούμε κάποιες σημαντικές μειώσεις στον αριθμό των διαδρομών με ταξί, χωρίς να υπάρχει σαφές μοτίβο για το πότε και γιατί συμβαίνουν. Επομένως, θα δούμε αν η μείωση των διαδρομών συσχετίζεται με τις αργίες. Ελέγχουμε αν υπάρχει συσχέτιση, συνδέοντας το σύνολο δεδομένων για τις διαδρομές ταξί της Νέας Υόρκης με το σύνολο δεδομένων για τις δημόσιες αργίες, με το ακόλουθο ερώτημα.

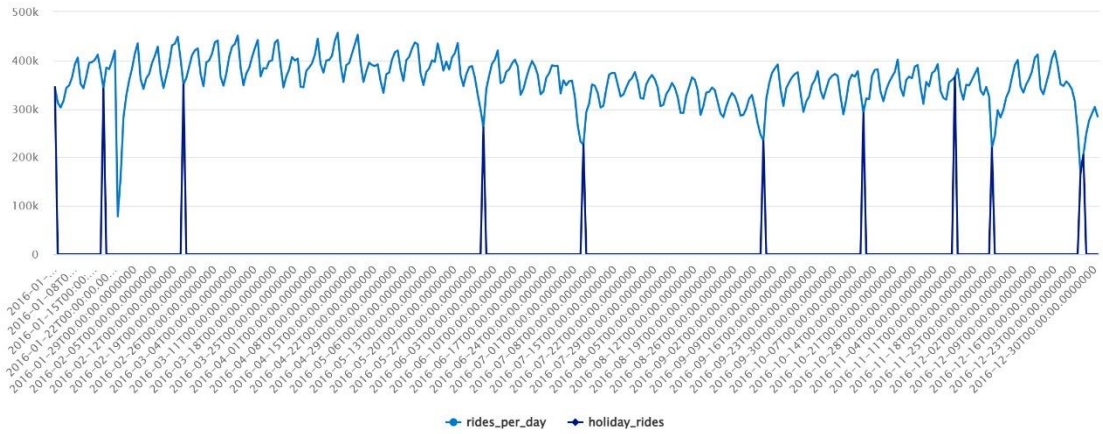


Στον πίνακα αποτελεσμάτων βλέπουμε ξεκάθαρα το σύνολο των διαδρομών που έγιναν κατά τις επίσημες αργίες του έτους 2016.



current_day	rides_per_day	holiday	date	holiday_rides
2016-01-01T00:00:00.0000000	345037	New Year's Day	2016-01-01T00:00:00.0000000	345037
2016-01-02T00:00:00.0000000	312831	(NULL)	(NULL)	0
2016-01-03T00:00:00.0000000	302878	(NULL)	(NULL)	0
2016-01-04T00:00:00.0000000	316171	(NULL)	(NULL)	0
2016-01-05T00:00:00.0000000	343251	(NULL)	(NULL)	0
2016-01-06T00:00:00.0000000	348516	(NULL)	(NULL)	0
2016-01-07T00:00:00.0000000	364894	(NULL)	(NULL)	0
2016-01-08T00:00:00.0000000	392070	(NULL)	(NULL)	0
2016-01-09T00:00:00.0000000	405825	(NULL)	(NULL)	0
2016-01-10T00:00:00.0000000	351788	(NULL)	(NULL)	0
2016-01-11T00:00:00.0000000	342651	(NULL)	(NULL)	0
2016-01-12T00:00:00.0000000	367390	(NULL)	(NULL)	0

Άρα, επισημαίνουμε τον αριθμό των διαδρομών ταξί κατά τη διάρκεια των αργιών επιλέγοντας «current_day» για τη στήλη «Category» και «rides_per_day», «holiday_rides» ως στήλες Legend (series). Επομένως, στο διάγραμμα που προκύπτει βλέπουμε τις διαδρομές ανά ημέρα του έτους 2016 αντιστοιχισμένες με τις αργίες του ίδιου έτους.



Αυτό που προκύπτει από το διάγραμμα είναι ότι κατά τη διάρκεια των αργιών ο αριθμός των διαδρομών με ταξί είναι χαμηλότερος σε σχέση με τις υπόλοιπες ημέρες. Υπάρχει ακόμα μια ανεξήγητη μεγάλη πτώση στις 23 Ιανουαρίου. Ας ελέγξουμε τον καιρό στη Νέα Υόρκη εκείνη την ημέρα, κάνοντας αναζήτηση στο σύνολο καιρικών δεδομένων (Weather Data).

SQL script 1

Run | Undo | Publish | Query plan | Connect to: Built-in | Use database: master

```
SELECT
    AVG(windspeed) AS avg_windspeed,
    MIN(windspeed) AS min_windspeed,
    MAX(windspeed) AS max_windspeed,
    AVG(temperature) AS avg_temperature,
    MIN(temperature) AS min_temperature,
    MAX(temperature) AS max_temperature,
    AVG(sealvlpressure) AS avg_sealvlpressure,
    MIN(sealvlpressure) AS min_sealvlpressure,
    MAX(sealvlpressure) AS max_sealvlpressure,
    AVG(precipdepth) AS avg_precipdepth,
    MIN(precipdepth) AS min_precipdepth,
    MAX(precipdepth) AS max_precipdepth,
    AVG(snowdepth) AS avg_snowdepth,
    MIN(snowdepth) AS min_snowdepth,
    MAX(snowdepth) AS max_snowdepth
FROM
    OPENROWSET(
        BULK 'https://azureopendatastorage.blob.core.windows.net/isdweatherdatacontainer/ISDWeather/year=*/month=*/*.parquet',
        FORMAT='PARQUET'
    ) AS [weather]
WHERE countryorregion = 'US' AND CAST([datetime] AS DATE) = '2016-01-23' AND stationname = 'JOHN F KENNEDY INTERNATIONAL AIRPORT'
```

Results | Messages

View: Table | Chart | Export results

Search

avg_windspeed	min_windspeed	max_windspeed	avg_temperatu...	min_temperat...	max_temperat...	avg_sealvlpres...	min_sealvlpres...	max_sealvlpres...	avg_precipdepth	min_precipdep...	max_precipde...
10.0595744680...	3.6	14.9	-1.8191489361...	-3	-0.6	1010.50625	1001.8	1022.6	17.4545454545...	0	70

Τα αποτελέσματα του ερωτήματος δείχνουν ότι η πτώση του αριθμού των διαδρομών με ταξί σημειώθηκε επειδή:

- Η χιονοθύελλα εκείνη την ημέρα στη Νέα Υόρκη ήταν ισχυρή (~30 cm).
- Έκανε κρύο (η θερμοκρασία ήταν κάτω από τους μηδέν βαθμούς Κελσίου).
- Είχε αέρα (~10 m/s).

Αυτό το σενάριο έδειξε πώς ένας αναλυτής δεδομένων μπορεί να εκτελέσει γρήγορα διερευνητική ανάλυση δεδομένων. Μπορείτε να συνδυάσετε διαφορετικά σύνολα δεδομένων χρησιμοποιώντας serverless SQL pool και να απεικονίσετε τα αποτελέσματα χρησιμοποιώντας το Azure Synapse Studio.

Σημειώνεται πως η προεπιλεγμένη ταξινόμηση είναι SQL_Latin1_General_CP1_CI_ASIf. Για μια μη προεπιλεγμένη ταξινόμηση, λάβετε υπόψη την ευαισθησία στην πεζότητα. Εάν δημιουργήσετε μια βάση δεδομένων με ευαισθησία στη γραφή των πεζών χαρακτήρων όταν καθορίζετε στήλες, βεβαιωθείτε ότι χρησιμοποιείτε το σωστό όνομα της στήλης. Ένα όνομα στήλης tpepPickupDateTime θα ήταν σωστό, ενώ το tpeppickupdatetime δεν θα λειτουργούσε σε μια μη προεπιλεγμένη ταξινόμηση.

4.2 Σενάριο 2ο | Εκπαίδευση ενός μοντέλου παλινδρόμησης με τη χρήση Azure Machine Learning

Στο συγκεκριμένο σενάριο θα δούμε πώς μπορούμε να εκπαιδεύσουμε και να αναπτύξουμε ένα μοντέλο παλινδρόμησης (Regression Model) χωρίς την χρήση κώδικα, χρησιμοποιώντας τον σχεδιαστή του Azure Machine Learning. Το Azure Machine Learning είναι μια υπηρεσία cloud για την επιτάχυνση και τη διαχείριση του κύκλου ζωής των έργων μηχανικής μάθησης (ML). Οι επαγγελματίες ML, οι επιστήμονες δεδομένων και οι μηχανικοί μπορούν να τη χρησιμοποιούν στις καθημερινές ροές εργασίας τους για την εκτίμηση μοντέλων και τη διαχείριση των λειτουργιών μηχανικής μάθησης (MLOps).

Προηγουμένως είδαμε πως η μηχανική μάθηση χρησιμοποιεί αλγόριθμους για τον εντοπισμό μοτίβων μέσα στα δεδομένα και αυτά τα μοτίβα στη συνέχεια χρησιμοποιούνται για τη δημιουργία ενός μοντέλου δεδομένων που μπορεί να κάνει προβλέψεις. Η παλινδρόμηση είναι μια μορφή μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη μιας διακριτής, συνήθως αριθμητικής, τιμής ή ετικέτας με βάση τα χαρακτηριστικά ενός στοιχείου. Για παράδειγμα, μια εταιρεία πωλήσεων αυτοκινήτων μπορεί να χρησιμοποιήσει τα χαρακτηριστικά ενός αυτοκινήτου, όπως το μέγεθος του κινητήρα, τον αριθμό θέσεων, τα χιλιόμετρα, και ούτω καθεξής για να προβλέψει την πιθανή τιμή πώλησης. Στην περίπτωση αυτή, τα χαρακτηριστικά του αυτοκινήτου είναι τα χαρακτηριστικά και η τιμή πώλησης είναι η ετικέτα που θα προβλέψουμε. Η παλινδρόμηση είναι ένα παράδειγμα εποπτευόμενης τεχνικής μηχανικής μάθησης (Supervised Machine Learning) στην οποία εκπαιδεύεται ένα μοντέλο χρησιμοποιώντας δεδομένα. Τα δεδομένα περιλαμβάνουν τόσο τα χαρακτηριστικά όσο και τις γνωστές τιμές για την ετικέτα, έτσι ώστε το μοντέλο να μάθει να προσαρμόζει τους συνδυασμούς χαρακτηριστικών στην ετικέτα. Στη συνέχεια, μετά την ολοκλήρωση της εκπαίδευσης, μπορούμε να χρησιμοποιήσουμε το μοντέλο για να προβλέψουμε τιμές για νέα αντικείμενα για τα οποία η ετικέτα είναι άγνωστη.

4.2.1: Δημιουργία ενός χώρου εργασίας Azure ML

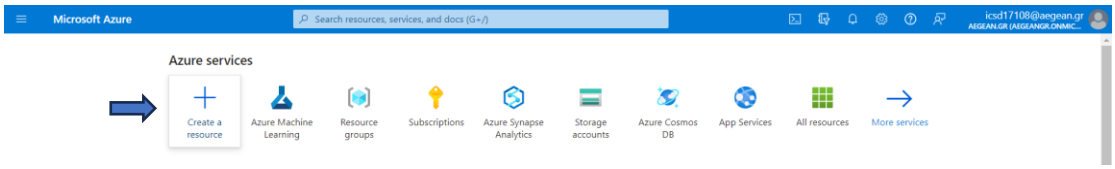
Προκειμένου να χρησιμοποιήσουμε το Azure ML, δημιουργούμε έναν χώρο εργασίας (workspace) στη συνδρομή μας στην πλατφόρμα. Στη συνέχεια, μπορούμε να χρησιμοποιήσουμε αυτόν τον χώρο εργασίας για τη διαχείριση δεδομένων, υπολογιστικών

πόρων, κώδικα, μοντέλων και άλλων αντικειμένων που σχετίζονται με τους φόρτους εργασίας μηχανικής μάθησης.

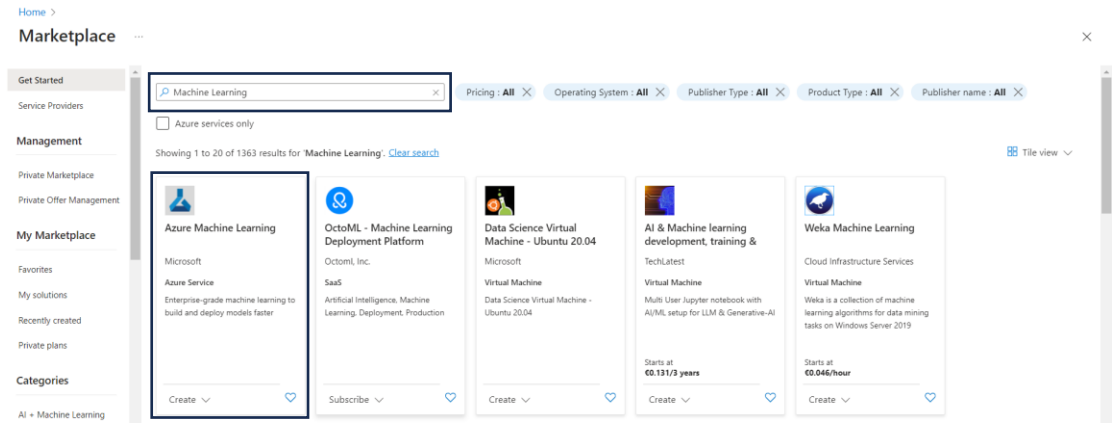
Σε περίπτωση που δεν διαθέτουμε ήδη ένα workspace, ακολουθούμε τα παρακάτω βήματα για να δημιουργήσουμε ένα χώρο εργασίας:

1. Συνδεόμαστε στην πύλη Microsoft Azure χρησιμοποιώντας τα διαπιστευτήριά μας.

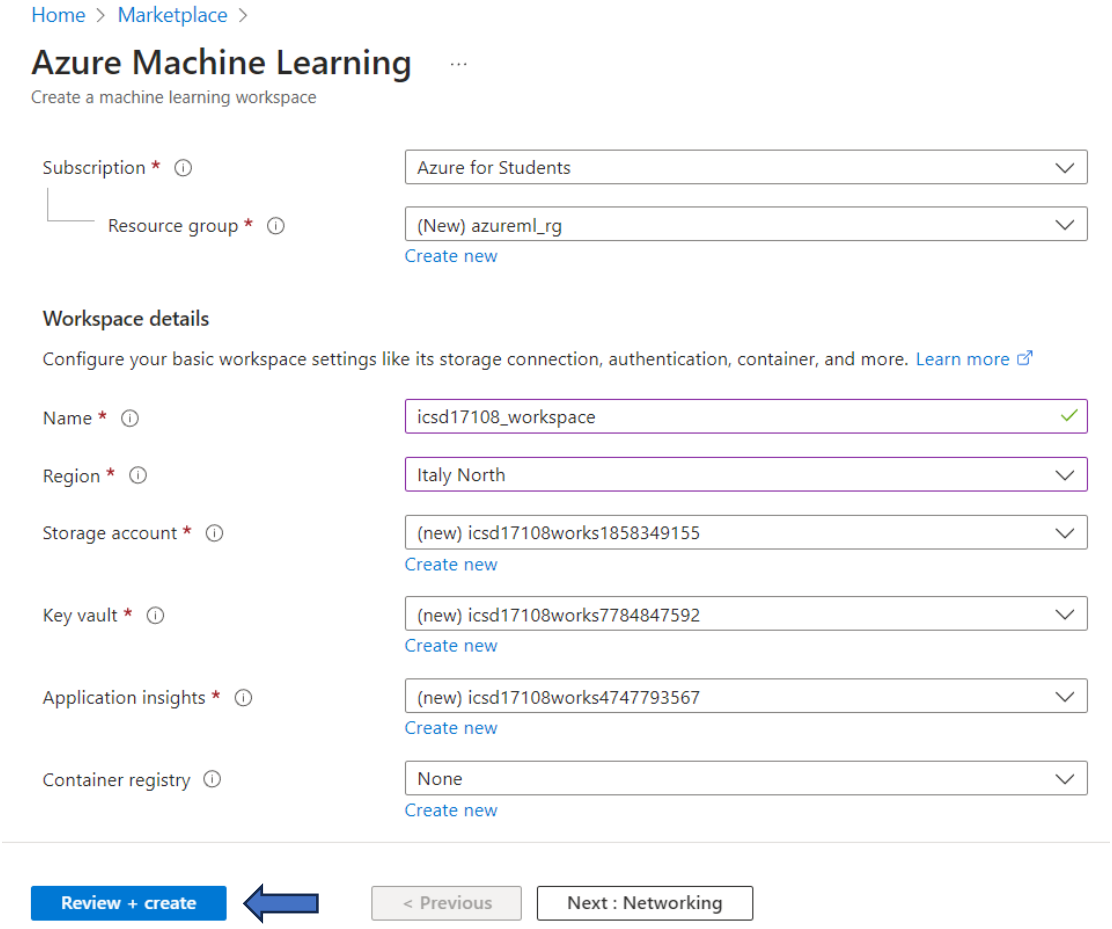
2. Επιλέγουμε + *Create a resource* (Δημιουργία πόρου).



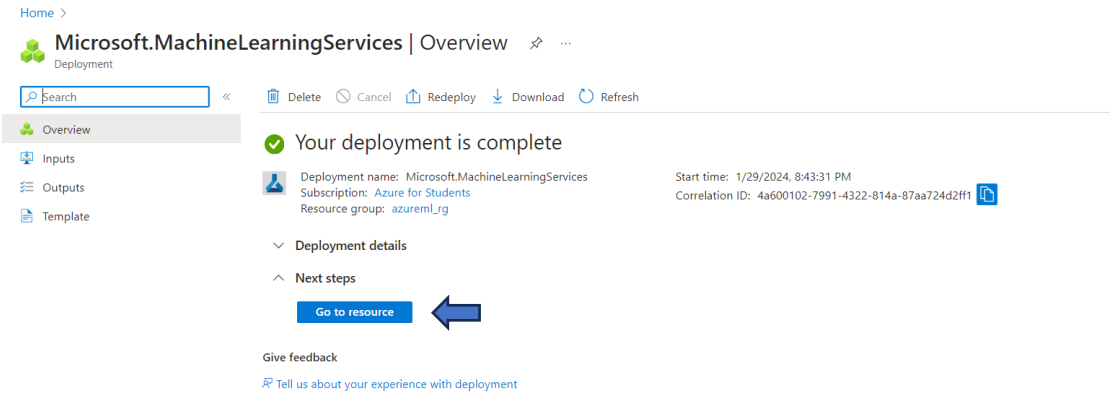
Έπειτα κάνουμε αναζήτηση Machine Learning και δημιουργούμε έναν νέο πόρο Azure Machine Learning με τις ακόλουθες ρυθμίσεις:



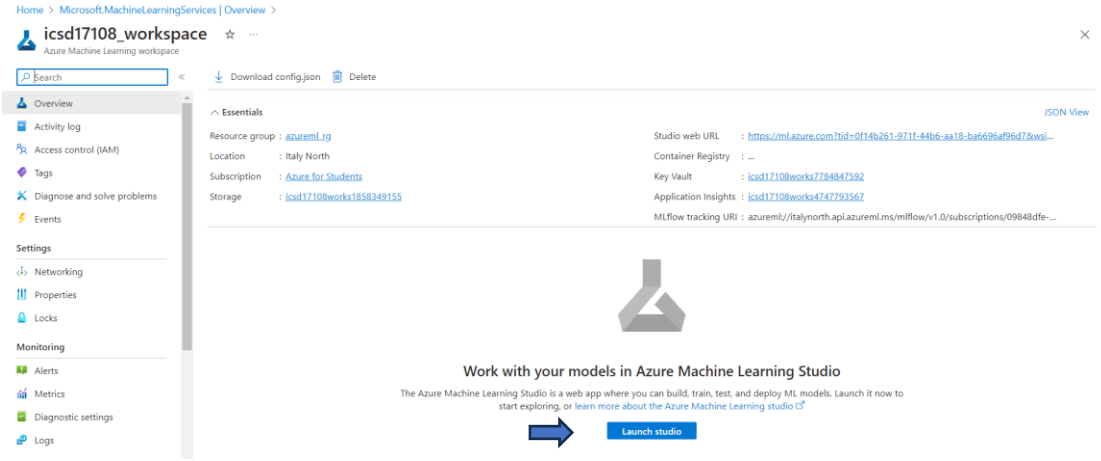
- **Subscription**: Η συνδρομή μας στο Azure.
- **Resource group**: Δημιουργούμε νέο ή επιλέγουμε μια ομάδα πόρων.
- **Name**: Εισάγουμε ένα μοναδικό όνομα για τον χώρο εργασίας μας.
- **Region**: Επιλέγουμε τη γεωγραφική περιοχή που βρίσκεται πιο κοντά σε εμάς.
- **Storage account**, **Key vault**, **Applications insights**: Συμπληρώνονται αυτόματα εφόσον ορίσουμε προηγουμένως το όνομα.



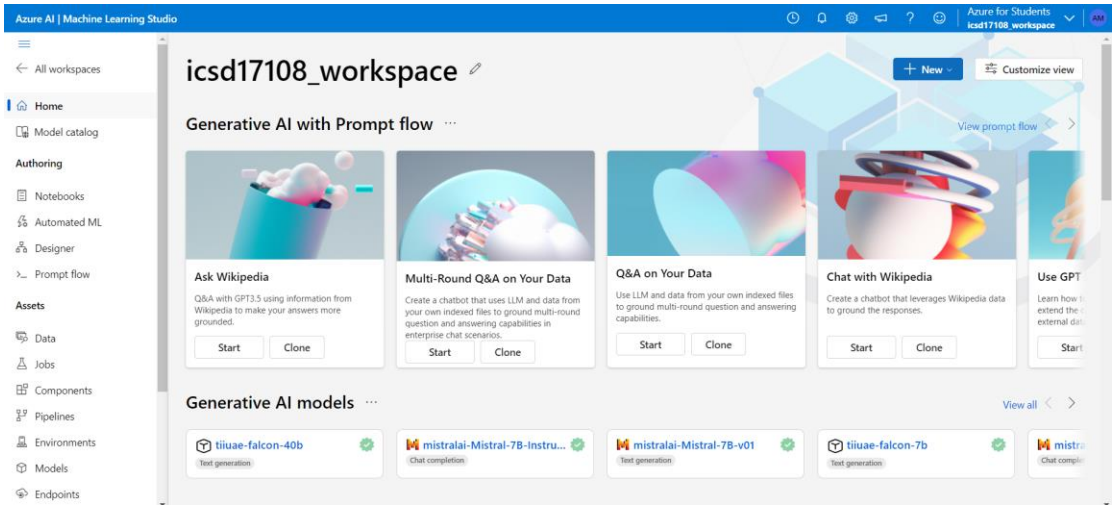
3. Εφόσον πατήσαμε *Review + create* και μετά ξανά *Create*, περιμένουμε μερικά λεπτά έως ότου δημιουργηθεί ο χώρος εργασίας μας. Θα ενημερωθούμε με σχετικό μήνυμα κατά την ολοκλήρωση της διαδικασίας.



4. Στη συνέχεια πατάμε *Go to resource* όπου μεταφερόμαστε στη σελίδα επισκόπησης (Overview) του χώρου εργασίας μας. Τώρα μπορούμε να εκκινήσουμε το Microsoft Azure Machine Learning Studio επιλέγοντας *Launch studio*.



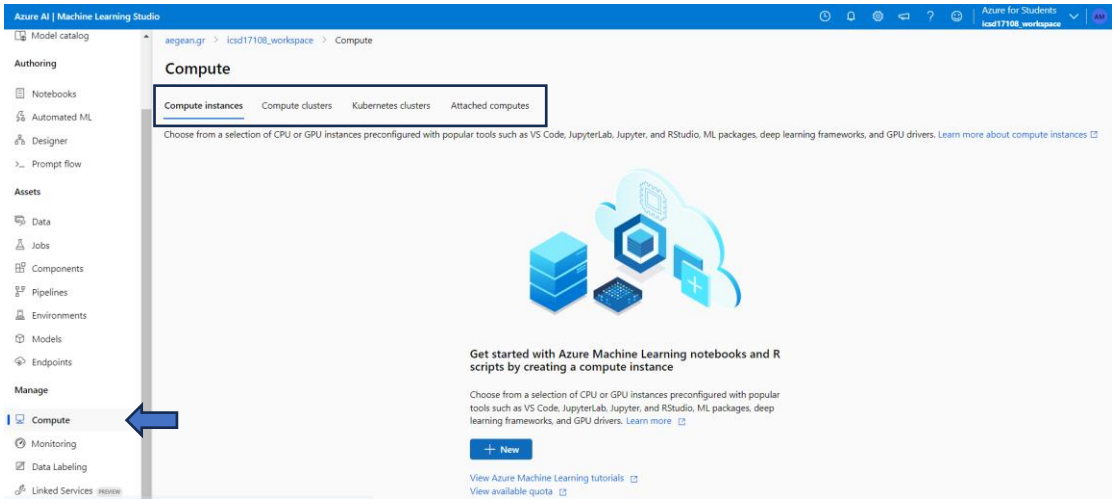
Εναλλακτικά ανοίγουμε μια νέα καρτέλα του προγράμματος περιήγησης και πληκτρολογούμε τη διεύθυνση *https://ml.azure.com*, όπου συνδεόμαστε στο Microsoft Azure Machine Learning Studio χρησιμοποιώντας το λογαριασμό μας στη Microsoft.



4.2.2: Δημιουργία υπολογιστικών πόρων

Οι υπολογιστικοί πόροι που θα χρησιμοποιήσουμε βασίζονται στο cloud, πάνω στους οποίους μπορούμε να εκτελέσουμε τις διαδικασίες εκπαίδευσης μοντέλων και διερεύνησης δεδομένων. Ακολουθούμε τα παρακάτω βήματα για την δημιουργία υπολογιστικών πόρων:

1. Στην πλατφόρμα επιλέγοντας το εικονίδιο ☰ στο επάνω αριστερό μέρος προβάλλονται οι διάφορες σελίδες του περιβάλλοντος εργασίας. Έπειτα, επιλέγουμε τη σελίδα *Compute* που βρίσκεται στην ομάδα *Manage* του μενού.

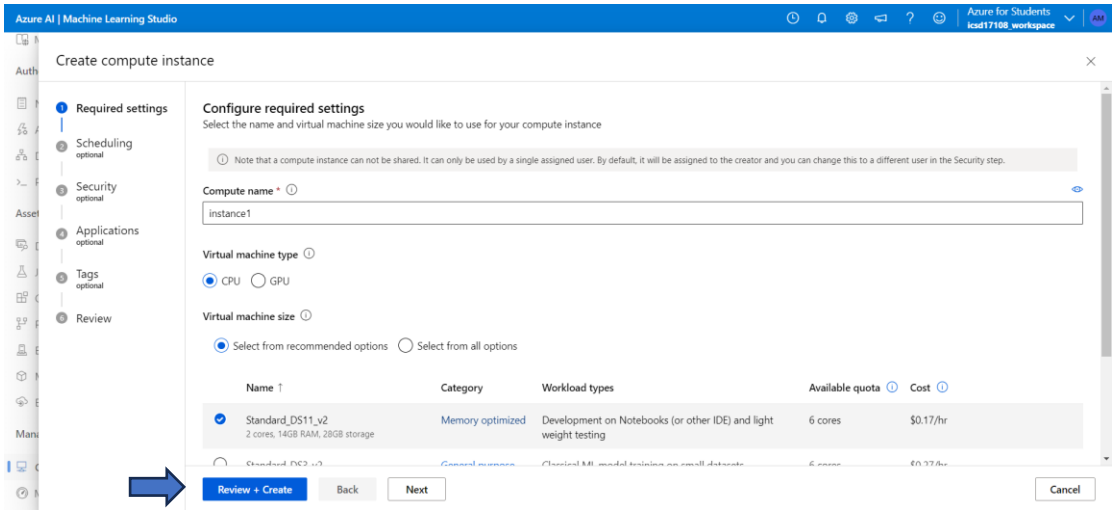


Στη συγκεκριμένη σελίδα διαχειριζόμαστε τους στόχους υπολογισμού για όλες τις δραστηριότητες ανάλυσης και διερεύνησης των δεδομένων μας. Υπάρχουν τέσσερα είδη υπολογιστικών πόρων που μπορούμε να δημιουργήσουμε:

- **Compute Instances (Μονάδες υπολογισμού)**: Οι σταθμοί εργασίας ή ανάπτυξης που μπορούν να χρησιμοποιούν οι επιστήμονες δεδομένων για να εργάζονται με δεδομένα και μοντέλα.
- **Compute Clusters (Συστάδες υπολογισμών)**: Οι επεκτάσιμες συστάδες εικονικών μηχανών για την κατά παραγγελία επεξεργασία πειραματικού κώδικα.
- **Inference Clusters (Συστάδες συμπερασμάτων)**: Οι στόχοι ανάπτυξης για τις υπηρεσίες πρόβλεψης που χρησιμοποιούν τα εκπαιδευμένα μοντέλα.
- **Attached Compute (Επισυναπτόμενος υπολογισμός)**: Οι συνδέσεις με υπάρχοντες υπολογιστικούς πόρους του Microsoft Azure, όπως εικονικές μηχανές ή Azure Databricks Clusters.

2. Στην καρτέλα *Compute Instances*, θα προσθέσουμε μια νέα μονάδα υπολογισμού, πατώντας *New*, με τις ακόλουθες ρυθμίσεις:

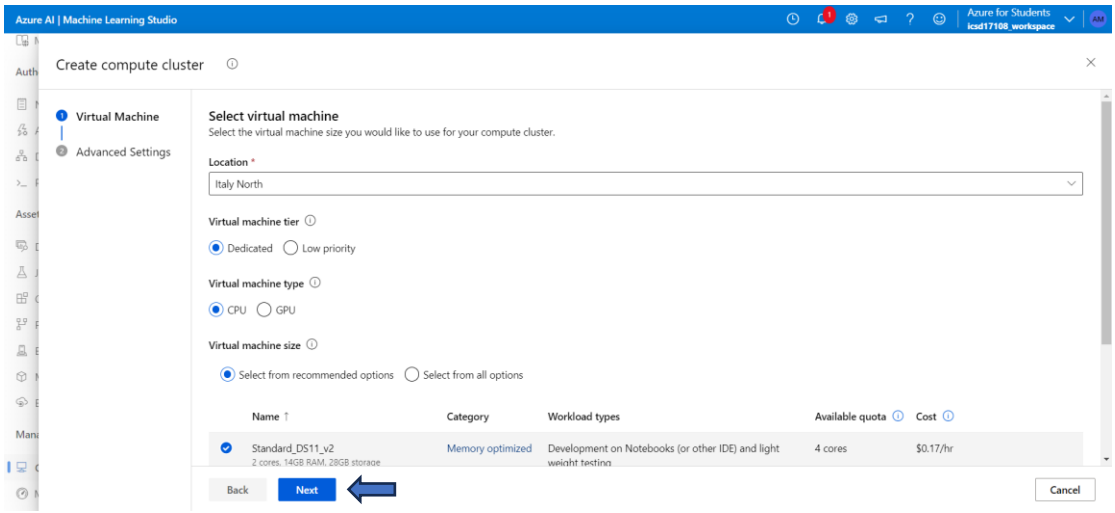
- **Compute name**: Εισάγουμε ένα μοναδικό όνομα.
- **Virtual machine type**: CPU
- **Virtual Machine size**: Standard_DS11_v2. Σε περίπτωση που δεν το βρίσκουμε επιλέγουμε *Select from all options* για να αναζητήσουμε και να επιλέξουμε αυτό το μέγεθος μηχανής.



3. Πατάμε Review + Create και μετά ξανά Create, αφού ελέγξουμε την ρύθμιση *Enable SSH Access: Unselected* στην καρτέλα *Security* ή στο τελικό review που θα αναγράφει *no*. Σημειώνεται πως την μονάδα υπολογισμού θα την χρησιμοποιήσουμε ως σταθμό εργασίας από τον οποίο θα δοκιμάσουμε και θα τεστάρουμε το μοντέλο μας.

4. Ενώ δημιουργείται η υπολογιστική μονάδα, πηγαίνουμε στην καρτέλα *Compute Clusters* και προσθέτουμε μια νέα συστάδα υπολογισμών με τις ακόλουθες ρυθμίσεις:

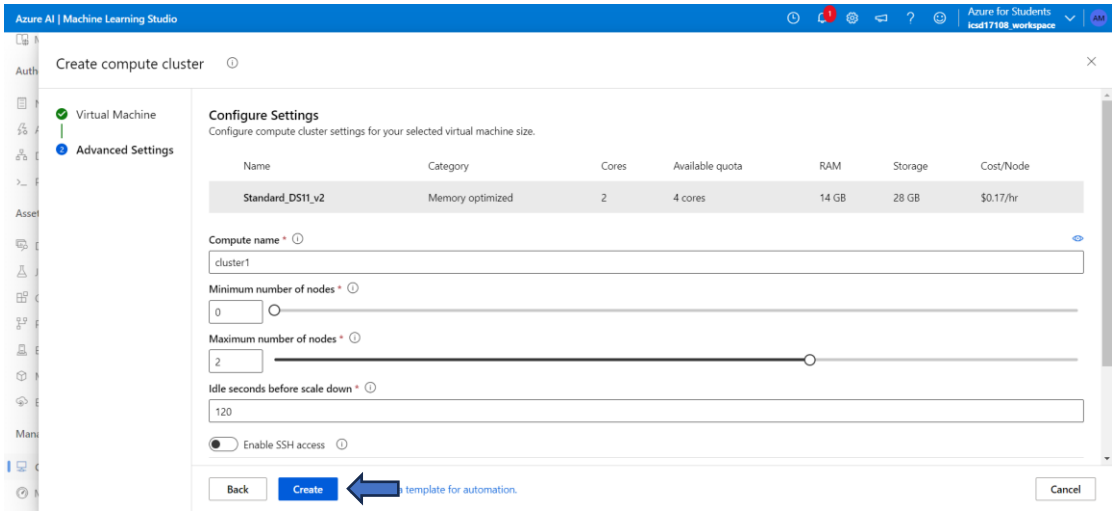
- **Virtual Machine tier:** Dedicated
- **Virtual Machine type:** CPU
- **Virtual Machine size**: Standard_DS11_v2. Σε περίπτωση που δεν το βρίσκουμε επιλέγουμε *Select from all options* για να αναζητήσουμε και να επιλέξουμε αυτό το μέγεθος μηχανής.



Στη συνέχεια πατάμε Next και πηγαίνουμε στην καρτέλα Advanced Settings.

- **Compute name**: Εισάγουμε ένα μοναδικό όνομα.
- **Minimum number of nodes**: 0

- **Maximum number of nodes**: 2
- **Idle seconds before scale down**: 120
- **Enable SSH access**: Unselected.



Τέλος, πατάμε *Create*. Σημειώνεται πως την συγκεκριμένη υπολογιστική μονάδα θα τη χρησιμοποιήσουμε για να εκπαιδεύσουμε ένα μοντέλο μηχανικής μάθησης.

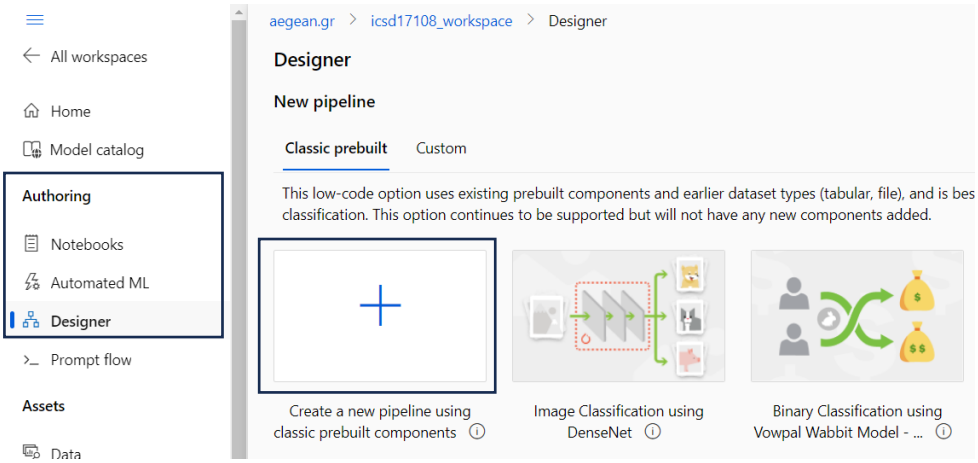
4.2.3: Εξερεύνηση δεδομένων

Προκειμένου να εκπαιδεύσουμε ένα μοντέλο παλινδρόμησης, χρειαζόμαστε ένα σύνολο δεδομένων που περιλαμβάνει ιστορικά χαρακτηριστικά, δηλαδή χαρακτηριστικά της οντότητας για την οποία θέλουμε να κάνουμε μία πρόβλεψη, και τις γνωστές τιμές ετικέτας, δηλαδή την αριθμητική τιμή που θέλουμε να προβλέψει το μοντέλο που θα εκπαιδεύσουμε.

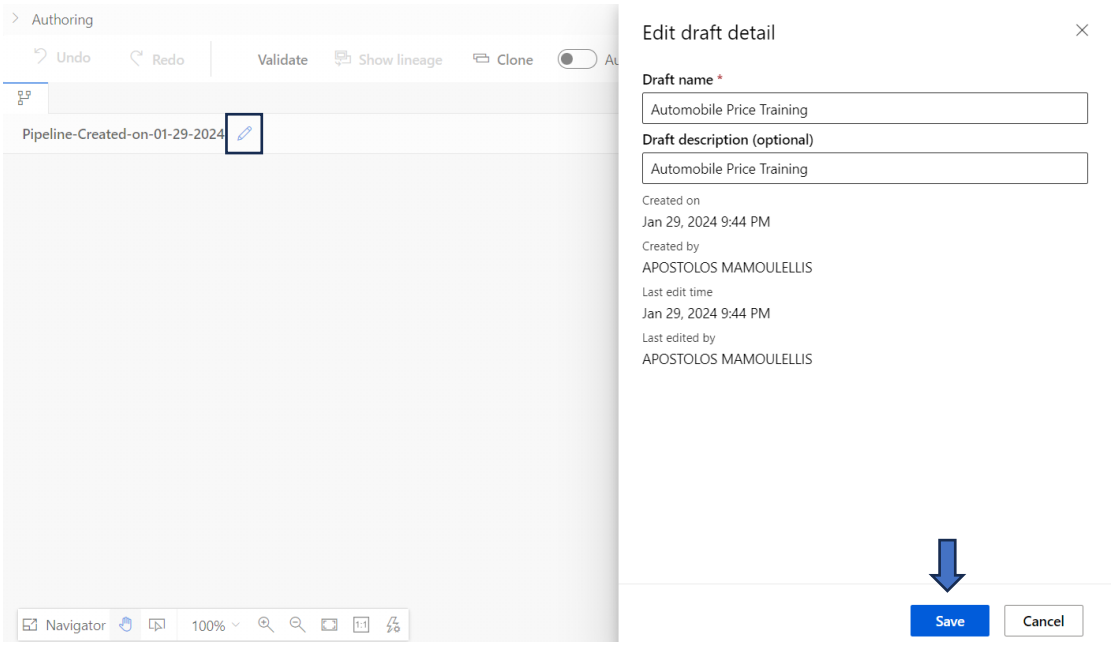
Δημιουργία ενός αγωγού (Pipeline)

Αρχικά, για να χρησιμοποιήσουμε τον σχεδιαστή του Microsoft Azure ML, δημιουργούμε έναν αγωγό για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Αυτός ο αγωγός ξεκινά με το σύνολο δεδομένων από το οποίο θέλουμε να εκπαιδεύσουμε το μοντέλο.

1. Στο μενού βλέπουμε τη σελίδα *Designer* στην ομάδα *Authoring* και επιλέγουμε το εικονίδιο + για να δημιουργήσουμε έναν νέο αγωγό.



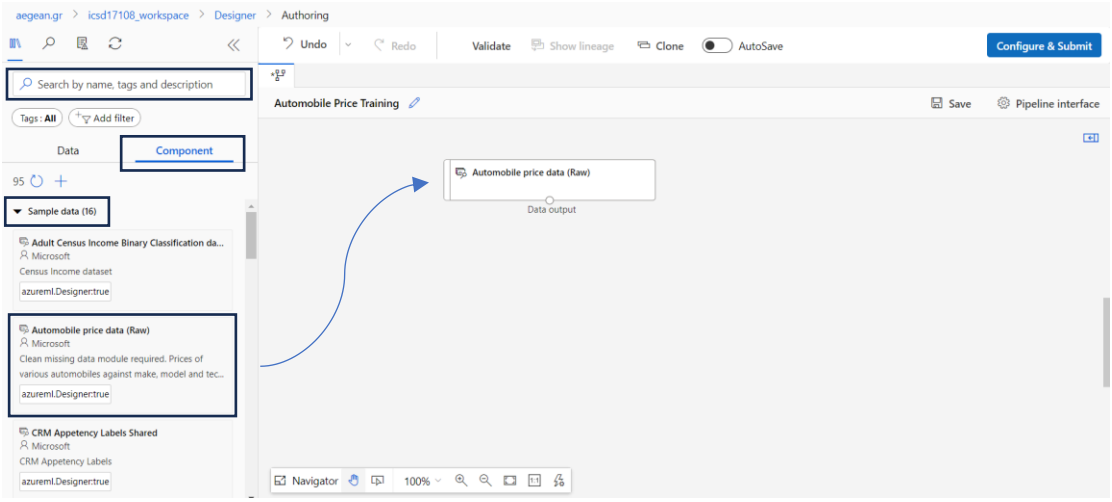
2. Στο παράθυρο κεντρικά αλλάζουμε το προεπιλεγμένο όνομα του αγωγού (*Pipeline-Created-on-date*) σε Auto Price Training, πατώντας το εικονίδιο Edit δίπλα στο όνομα του αγωγού.



Προσθήκη και εξερεύνηση συνόλου δεδομένων

Σε αυτό το σενάριο θα εκπαιδεύσουμε ένα μοντέλο παλινδρόμησης που προβλέπει την τιμή ενός αυτοκινήτου με βάση τα χαρακτηριστικά του. Το Azure Machine Learning περιλαμβάνει, ανάμεσα σε άλλα, ένα δείγμα συνόλου δεδομένων που μπορούμε να χρησιμοποιήσουμε για αυτό το μοντέλο.

1. Στην αριστερή πλευρά του σχεδιαστή, ανοίγουμε την ενότητα *Sample data*, όπου υπάρχουν όλα τα δείγματα συνόλων δεδομένων και σέρνουμε το σύνολο δεδομένων *Automobile price data (Raw)* από την ενότητα *Component* στον καμβά. Για μεγαλύτερη ευκολία κάνουμε αναζήτηση κι έπειτα το σέρνουμε με το ποντίκι το αποτέλεσμα στον καμβά.

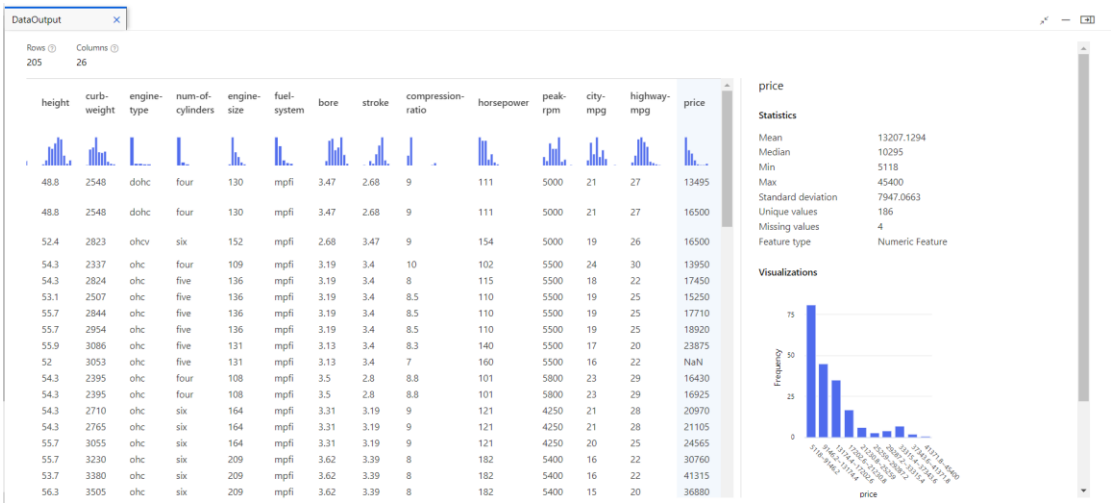


2. Κάνουμε δεξί κλικ στο σύνολο δεδομένων *Automobile price data (Raw)* στον καμβά και έπειτα στο μενού επιλέγουμε *Preview Data*.

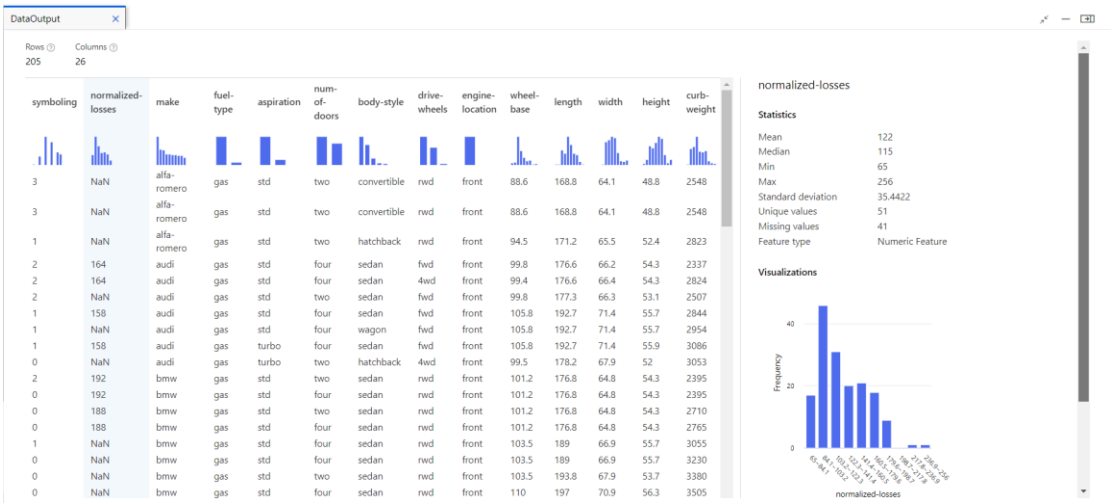
3. Επανεξετάζουμε το σχήμα των δεδομένων, σημειώνοντας ότι μπορούμε να δούμε τις κατανομές των διαφόρων στηλών ως ιστογράμματα.

4. Μετακινούμαστε προς τα δεξιά του συνόλου δεδομένων μέχρι να δούμε τη στήλη Price, καθώς αυτή είναι η ετικέτα που θα προβλέψει το μοντέλο μας.

5. Επιλέγουμε την επικεφαλίδα στήλης για τη στήλη price και βλέπουμε τις λεπτομέρειες που εμφανίζονται στο παράθυρο στα δεξιά. Αυτές περιλαμβάνουν διάφορα στατιστικά στοιχεία για τις τιμές της στήλης και ένα ιστόγραμμα που δείχνει την κατανομή των τιμών της στήλης.



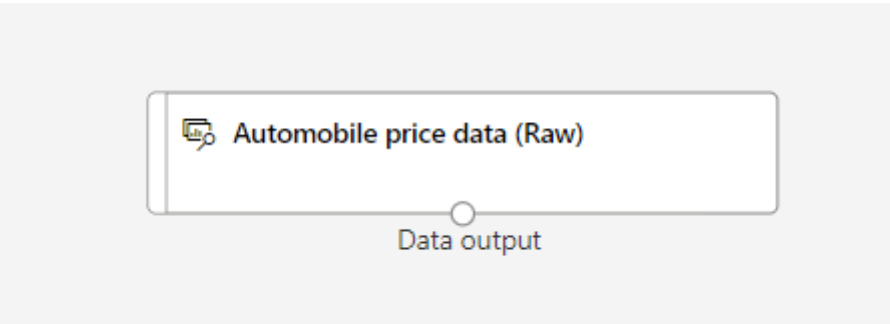
6. Κάνουμε κύλιση προς τα αριστερά και επιλέγουμε την επικεφαλίδα της στήλης normalized-losses. Στη συνέχεια, επανεξετάζουμε τα στατιστικά στοιχεία για αυτή τη στήλη, σημειώνοντας ότι υπάρχουν αρκετές ελλείπουσες τιμές σε αυτή τη στήλη. Αυτό θα περιορίσει τη χρησιμότητά της στην πρόβλεψη της ετικέτας τιμής και επομένως θέλουμε να την αποκλείσουμε από την εκτίμηση.



7. Παρατηρούμε τα στατιστικά στοιχεία για τις στήλες bore (η οπή ή διάμετρος κυλίνδρου σε έναν εμβολοφόρο κινητήρα), stroke (μια φάση του κύκλου του κινητήρα όπως διαδρομή συμπίεσης, διαδρομή εξαγωγής, κατά την οποία το έμβολο κινείται από πάνω προς τα κάτω ή αντίστροφα) και ιπποδύναμη (horsepower), σημειώνοντας τον αριθμό των τιμών που λείπουν. Αυτές οι στήλες έχουν σημαντικά λιγότερες ελλείπουσες τιμές από ό,τι οι κανονικοποιημένες απώλειες, οπότε μπορεί να εξακολουθούν να είναι χρήσιμες για την πρόβλεψη της τιμής, αν εξαιρέσουμε τις γραμμές όπου οι τιμές λείπουν από την εκτίμηση.

8. Συγκρίνουμε τις τιμές στις στήλες stroke, peak-rpm (μέγιστες στροφές ανά λεπτό) και city-mpg. Όλα αυτά μετρούνται σε διαφορετικές κλίμακες και είναι πιθανό οι μεγαλύτερες τιμές για τις peak-rpm να προκαλούν μεροληψία στον αλγόριθμο εκπαίδευσης και να δημιουργούν υπερβολική εξάρτηση από αυτή τη στήλη σε σύγκριση με στήλες με χαμηλότερες τιμές, όπως το stroke. Συνήθως, οι επιστήμονες δεδομένων μετριάζουν αυτή την πιθανή μεροληψία κανονικοποιώντας τις αριθμητικές στήλες ώστε να βρίσκονται σε παρόμοιες κλίμακες.

9. Κλείνουμε το παράθυρο απεικόνισης αποτελεσμάτων Automobile price data (Raw), ώστε να μπορέσουμε να δούμε το σύνολο δεδομένων στον καμβά ως εξής:



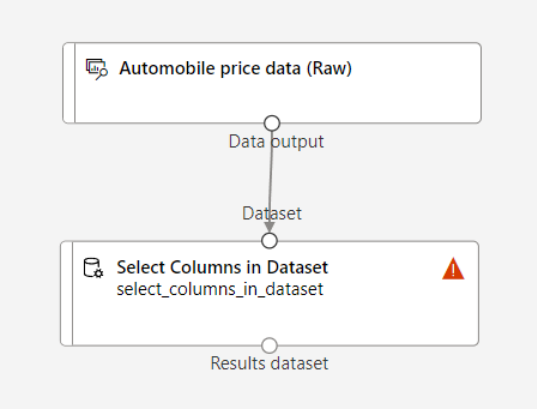
Προσθήκη μετασχηματισμών δεδομένων

Συνήθως εφαρμόζουμε μετασχηματισμούς δεδομένων για να προετοιμάσουμε τα δεδομένα για μοντελοποίηση. Στην περίπτωση των δεδομένων για τις τιμές των αυτοκινήτων, θα

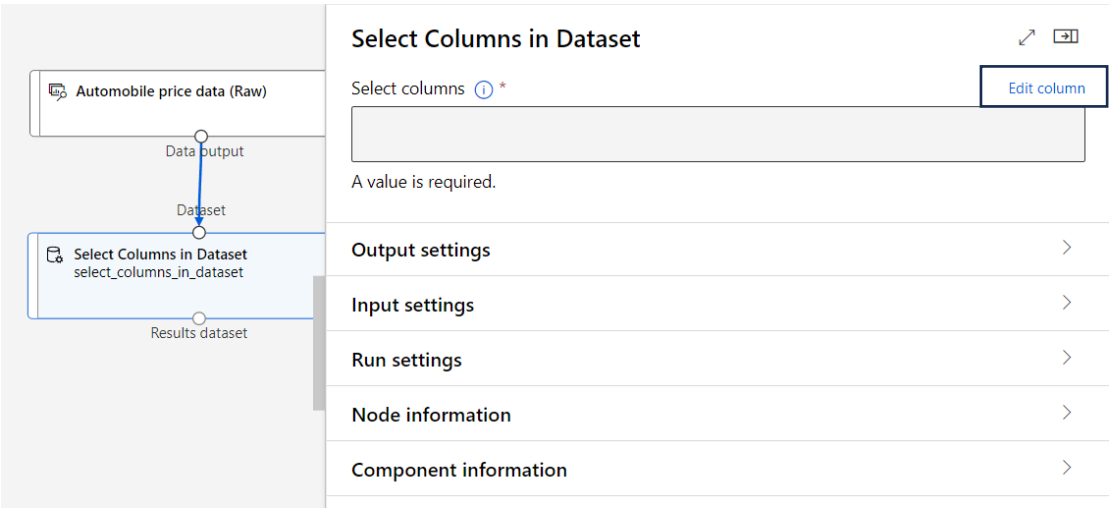
προσθέσουμε μετασχηματισμούς για να αντιμετωπίσουμε τα ζητήματα που εντοπίσαμε κατά τη διερεύνηση των δεδομένων.

1. Στο παράθυρο στα αριστερά, αναπτύσσουμε την ενότητα *Data Transformation*, η οποία περιέχει ένα ευρύ φάσμα ενοτήτων που μπορούμε να χρησιμοποιήσουμε για να μετασχηματίσουμε δεδομένα πριν από την εκπαίδευση του μοντέλου.

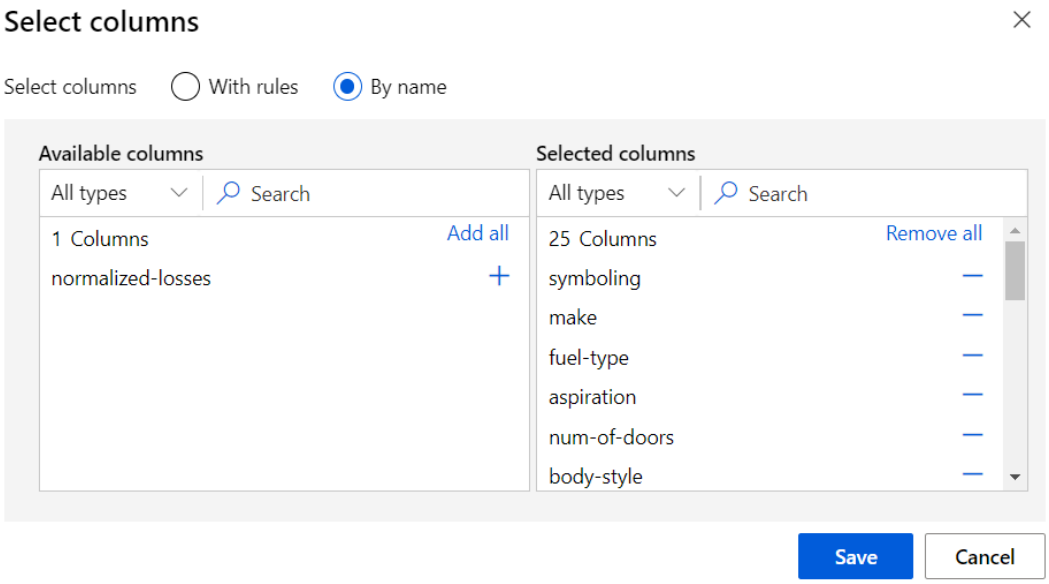
2. Σέρνουμε την ενότητα *Select Columns in Dataset* στον καμβά, κάτω από την ενότητα *Automobile price data (Raw)*. Στη συνέχεια, συνδέουμε την έξοδο στο κάτω μέρος της ενότητας *Automobile price data (Raw)* με την είσοδο στο πάνω μέρος της ενότητας *Select Columns in Dataset*, ως εξής:



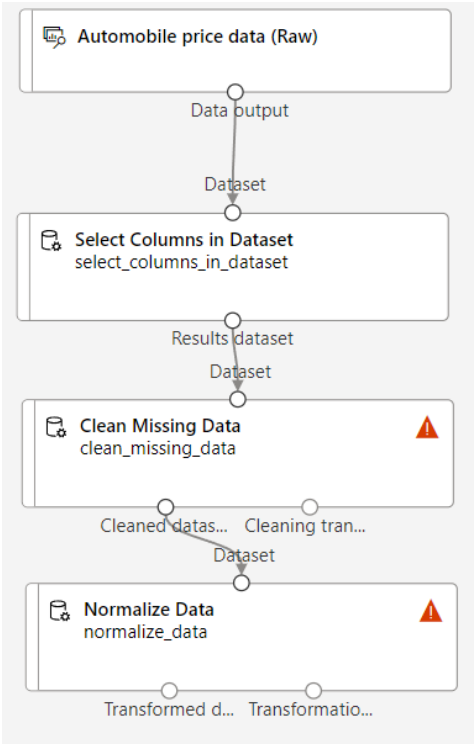
3. Επιλέγουμε την ενότητα με διπλό κλικ *Select Columns in Dataset* (Επιλογή στηλών στο σύνολο δεδομένων) και στο παράθυρο *Settings* (Ρυθμίσεις) στα δεξιά της, επιλέξτε *Edit column* (Επεξεργασία στήλης).



Στη συνέχεια, στο παράθυρο *Select columns* (Επιλογή στηλών), επιλέγουμε *By name* (Με όνομα) και χρησιμοποιούμε τους συνδέσμους + για να προσθέσουμε όλες τις στήλες εκτός από τις κανονικοποιημένες απώλειες, όπως εδώ:

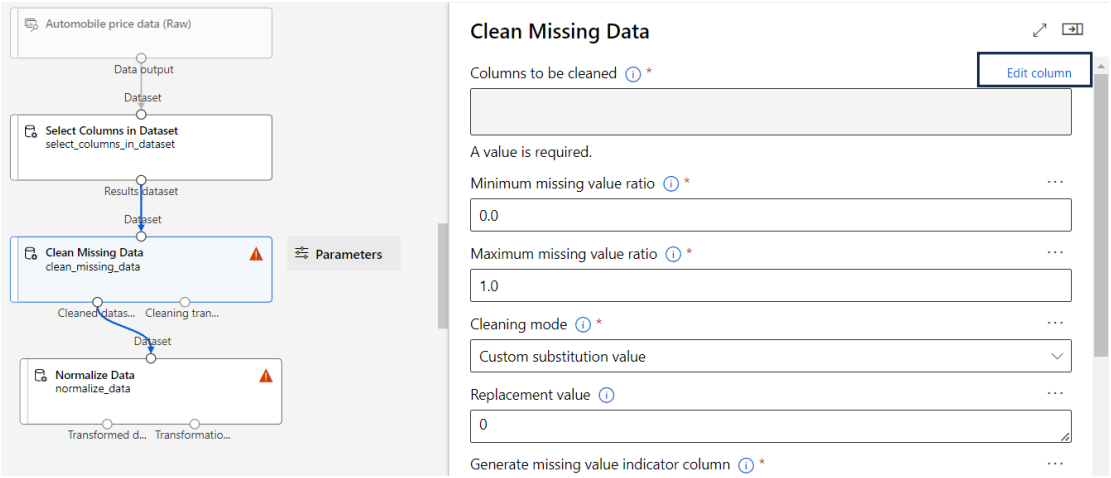


Ακολουθούμε τα υπόλοιπα βήματα, χρησιμοποιώντας την παρακάτω εικόνα ως αναφορά καθώς προσθέτουμε και ρυθμίζουμε τις απαιτούμενες ενότητες. Έπειτα, θα χτίσουμε τον αγωγό κάπως έτσι:

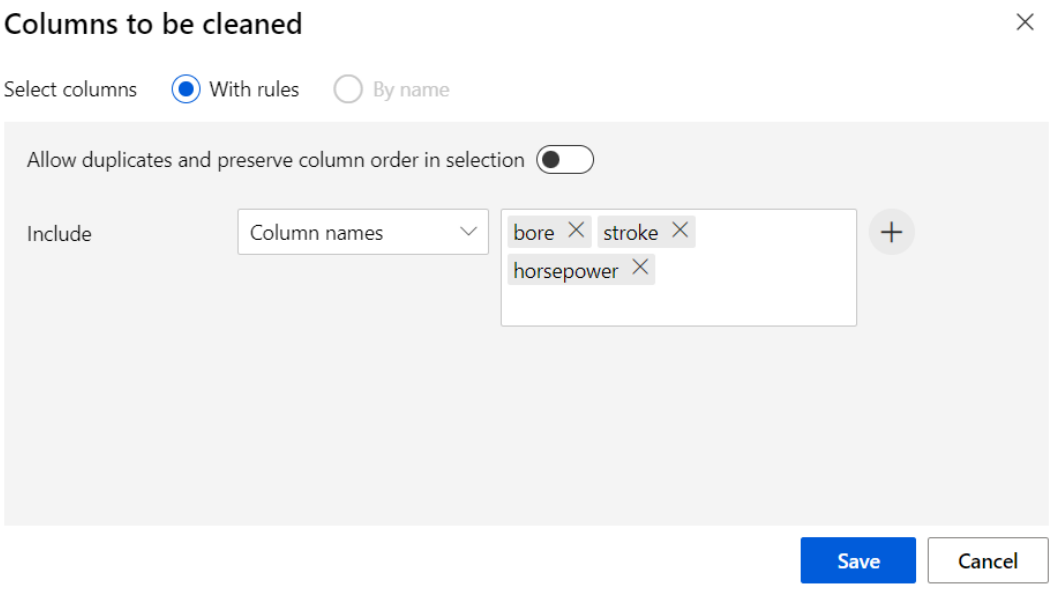


4. Σέρνουμε την ενότητα *Clean Missing Data* από την ενότητα *Data Transformations* (Μετασχηματισμοί δεδομένων) και την τοποθετούμε κάτω από την ενότητα *Select Columns in Dataset*. Στη συνέχεια, συνδέουμε την αριστερή έξοδο από την ενότητα *Select Columns in Dataset* στην είσοδο της ενότητας *Clean Missing Data*.

5. Επιλέγουμε την ενότητα *Clean Missing Data* με διπλό κλικ και στο παράθυρο ρυθμίσεων στα δεξιά, κάνουμε κλικ στην επιλογή *Edit column*.



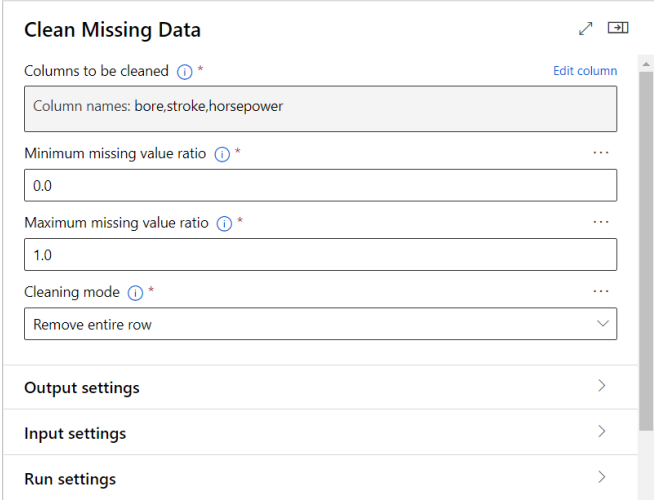
Στη συνέχεια, στο παράθυρο *Select columns*, επιλέγουμε *With rules*, στη λίστα *Include* επιλέγουμε *Column names* (Ονόματα στηλών) και στο πλαίσιο των ονομάτων στηλών πληκτρολογούμε *bore*, *stroke* και *horsepower*, προσέχοντας να ταιριάζει ακριβώς η ορθογραφία και η κεφαλαιοποίηση, όπως παρακάτω:



6. Με την ενότητα *Clean Missing Data* ακόμα επιλεγμένη, στο παράθυρο ρυθμίσεων, ορίζουμε τις ακόλουθες ρυθμίσεις διαμόρφωσης:

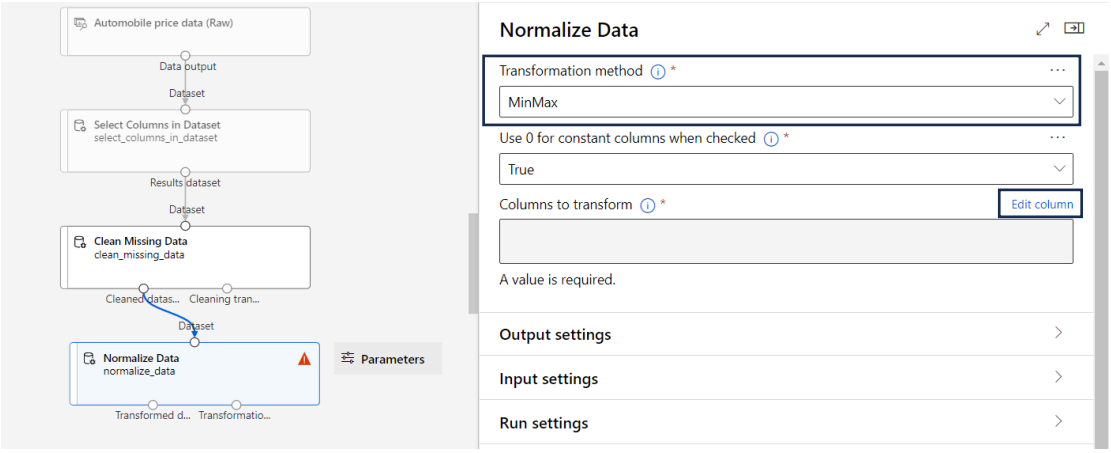
- **Minimum missing value ratio**: 0,0

- **Maximum missing value ratio**: 1,0
- **Cleaning mode**: Remove entire row.

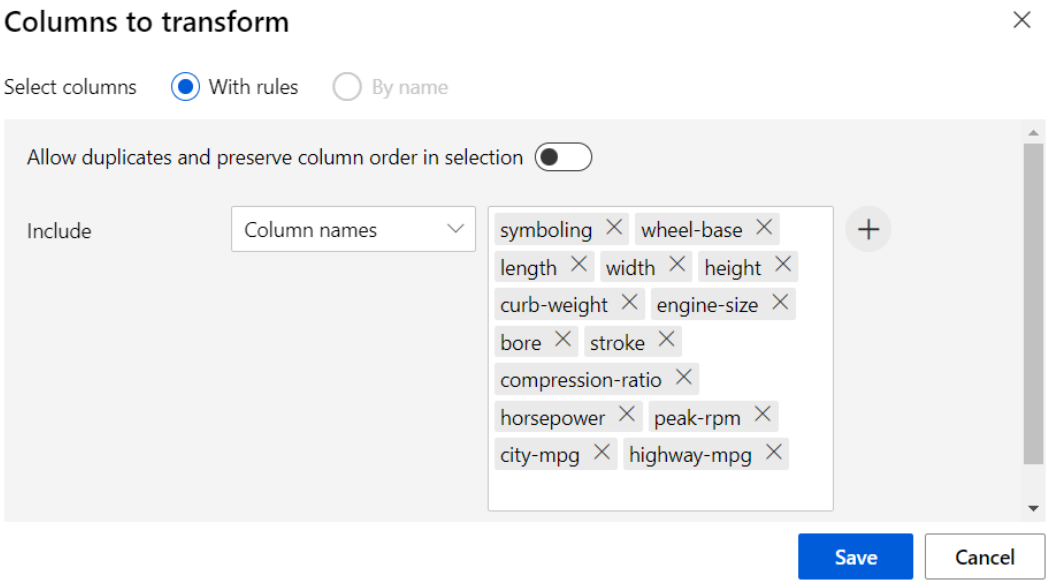


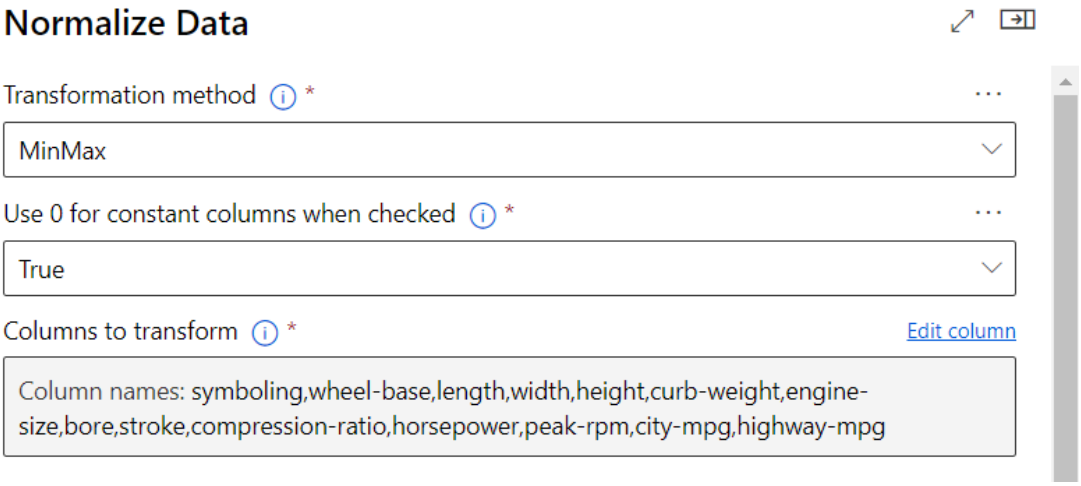
7. Σέρνουμε την ενότητα *Normalize Data* (Κανονικοποίηση δεδομένων) στον καμβά, κάτω από την ενότητα *Clean Missing Data*. Στη συνέχεια, συνδέουμε την αριστερότερη έξοδο από τη μονάδα *Clean Missing Data* στην είσοδο της μονάδας *Normalize Data*.

8. Επιλέγουμε την ενότητα *Normalize Data* με διπλό κλικ και βλέπουμε τις ρυθμίσεις της, σημειώνοντας ότι απαιτεί να καθορίσουμε τη μέθοδο μετασχηματισμού και τις στήλες που πρόκειται να μετασχηματιστούν. Επομένως, ορίζουμε τον μετασχηματισμό σε MinMax και επεξεργαζόμαστε τις στήλες (επιλέγουμε πάλι Edit column στο αντίστοιχο πεδίο) εφαρμόζοντας έναν κανόνα για να συμπεριλάβουμε τα ακόλουθα ονόματα στηλών, διασφαλίζοντας παράλληλα ότι ταιριάζουν ακριβώς με την ορθογραφία, την κεφαλαιοποίηση και την παύλα:



- symboling, wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg

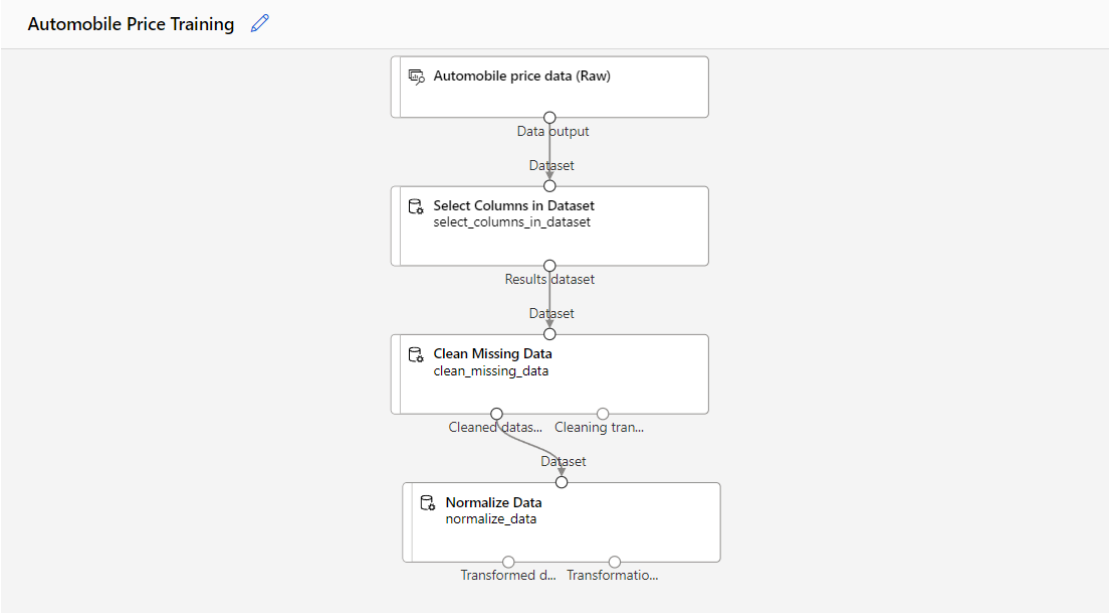




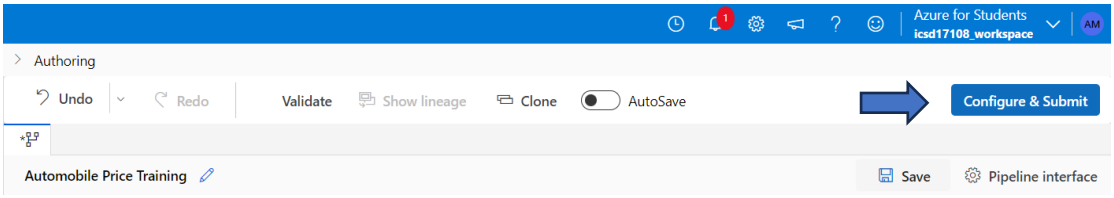
Εκτέλεση του αγωγού

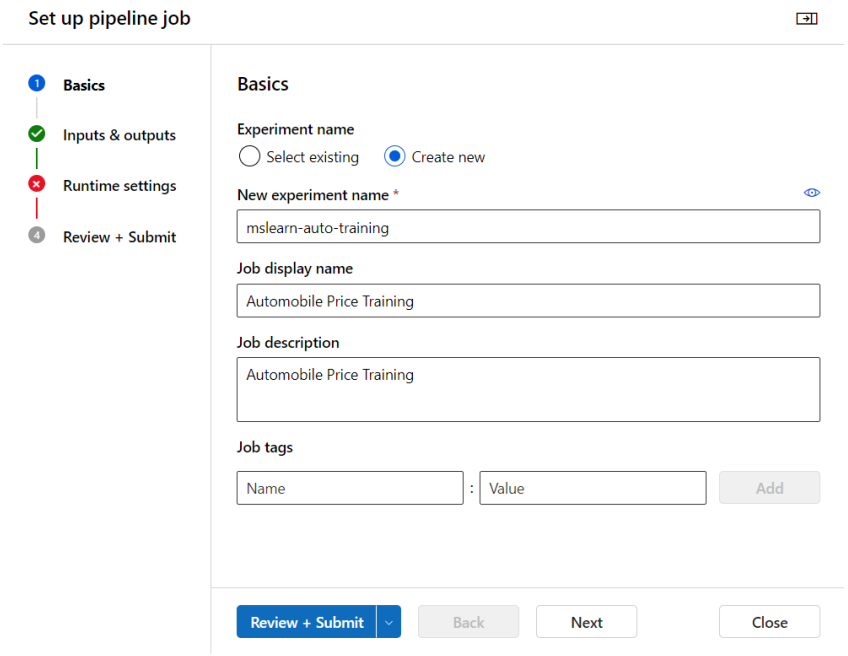
Για να εφαρμόσουμε τους μετασχηματισμούς των δεδομένων μας, πρέπει να εκτελέσουμε τον αγωγό ως πείραμα, επομένως ακολουθούμε τα παρακάτω βήματα:

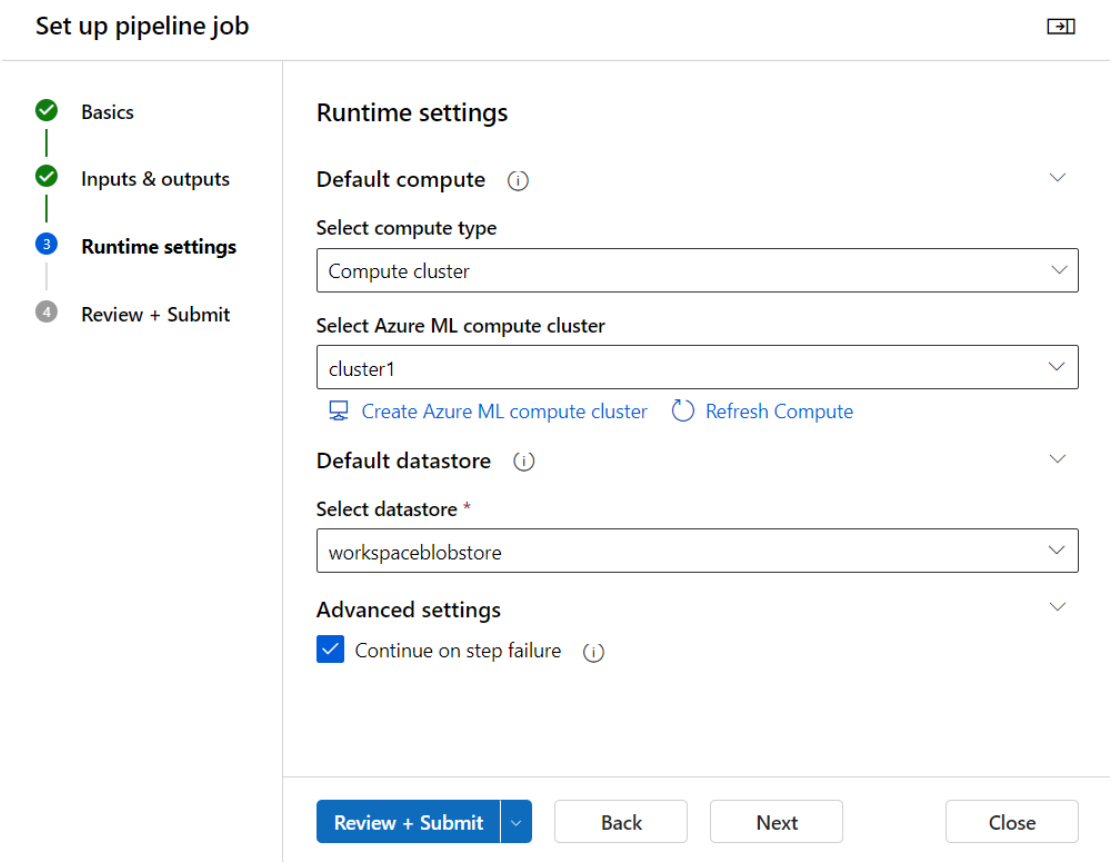
1. Επικυρώνουμε ότι ο αγωγός μας μοιάζει με αυτό:



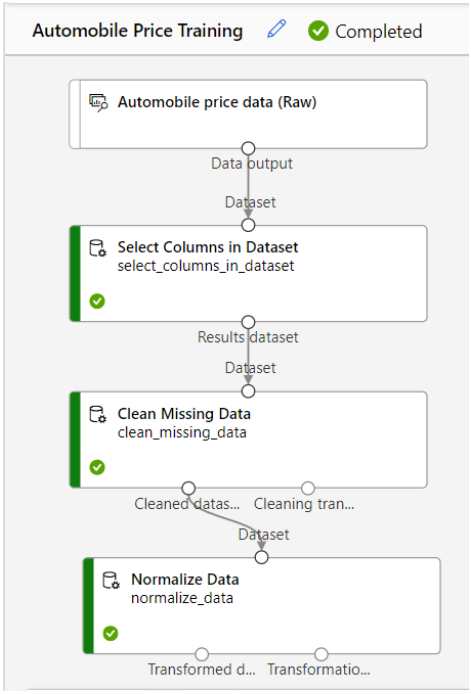
2. Επιλέγουμε *Configure & Submit* και εκτελούμε τον αγωγό ως νέο πείραμα με όνομα mslearn-auto-training στη συστάδα υπολογιστών μας.







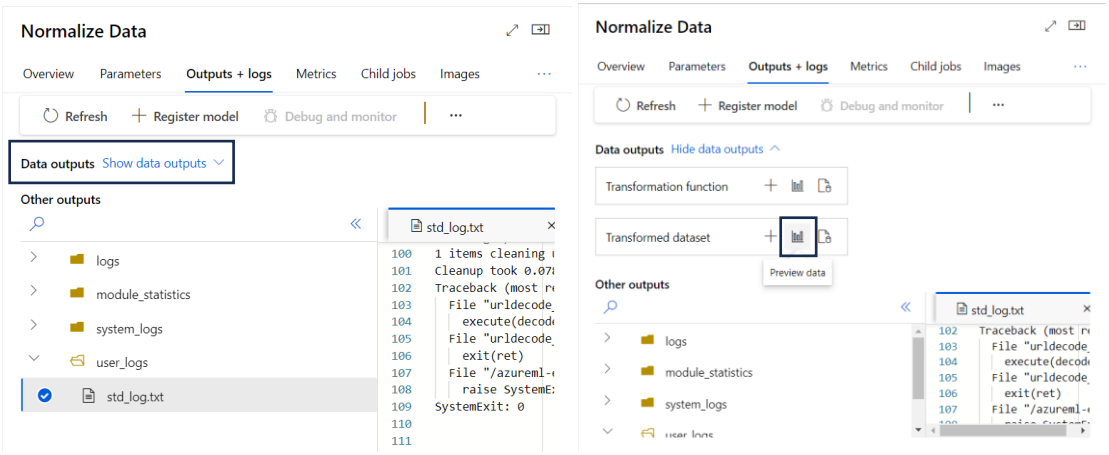
3. Τέλος, πατάμε *Review & Submit*, έπειτα πάλι *Submit* και περιμένουμε να ολοκληρωθεί η εκτέλεση. Αυτό μπορεί να διαρκέσει 5 λεπτά ή και περισσότερο. Όταν ολοκληρωθεί η εκτέλεση, οι ενότητες θα πρέπει να έχουν την εξής μορφή:



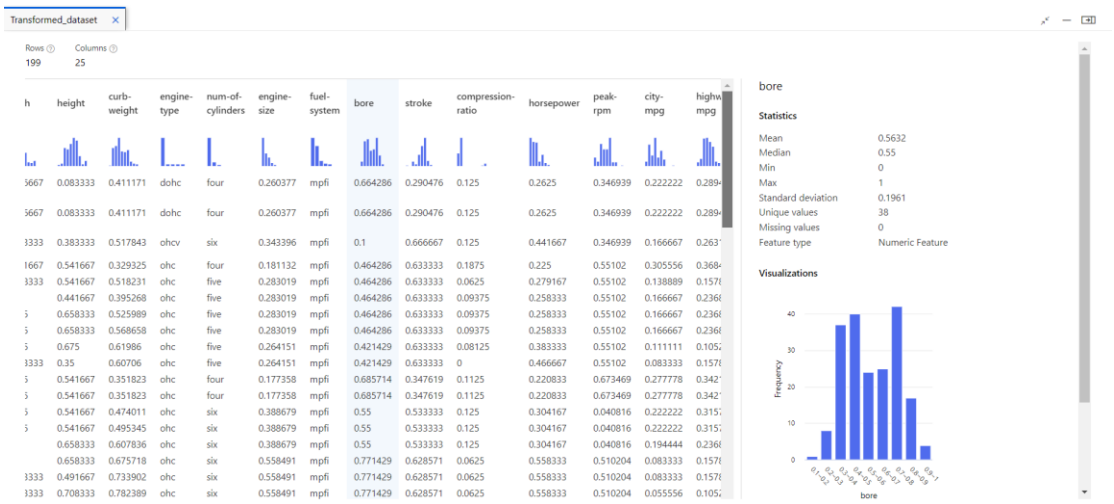
Προβολή των μετασχηματισμένων δεδομένων

Το σύνολο δεδομένων είναι τώρα έτοιμο για την εκπαίδευση του μοντέλου και για να το δούμε ακολουθούμε τα εξής βήματα:

1. Επιλέγουμε με διπλό κλικ την ολοκληρωμένη ενότητα *Normalize Data* και στο παράθυρο ρυθμίσεων στα δεξιά της, στην καρτέλα *Outputs + logs* (Έξοδοι + αρχεία καταγραφής), επιλέγουμε Show data outputs και Preview data.



2. Προβάλουμε τα δεδομένα, παρατηρώντας ότι η στήλη normalized-losses έχει αφαιρεθεί, όλες οι γραμμές περιέχουν δεδομένα για την οπή, τη διαδρομή και την ιπποδύναμη και οι αριθμητικές στήλες που επιλέξαμε έχουν κανονικοποιηθεί σε κοινή κλίμακα.

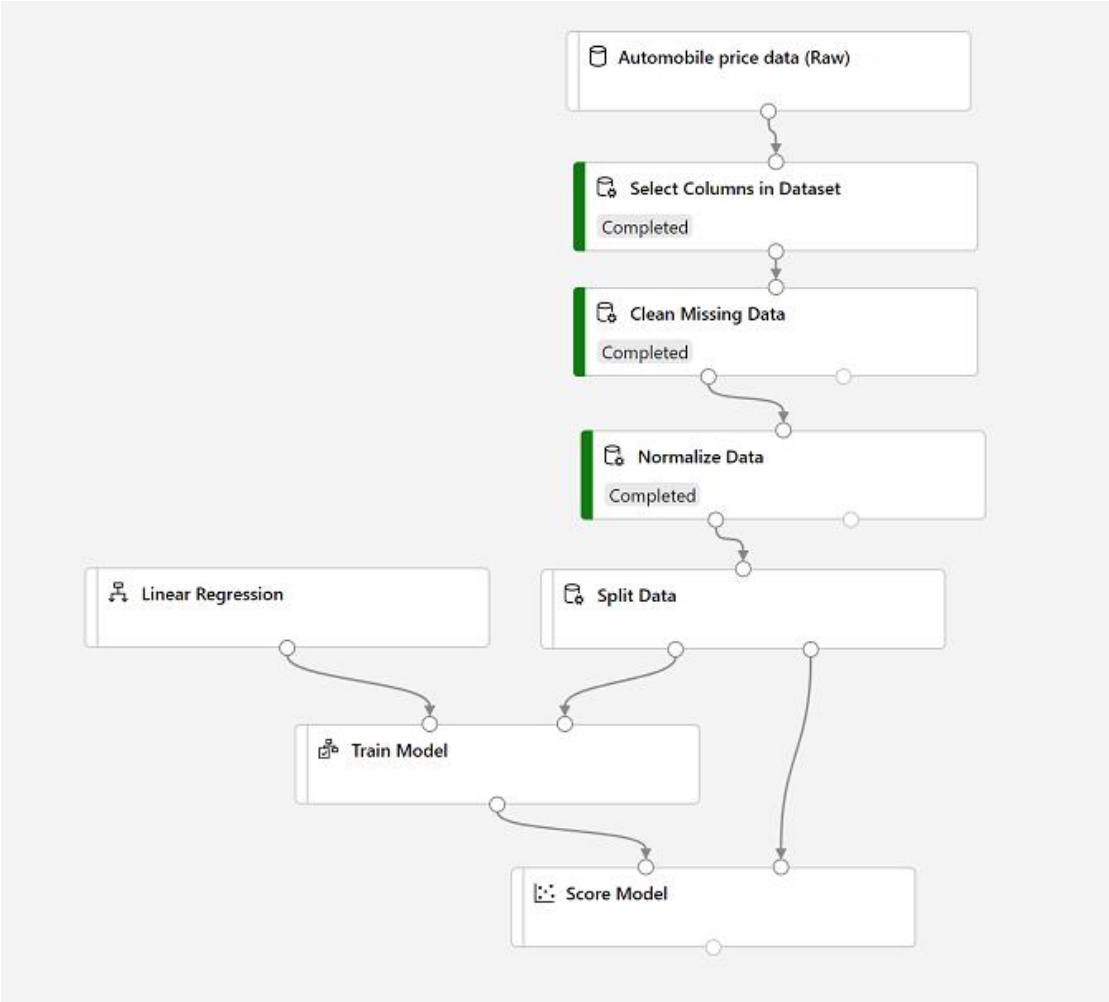


3. Κλείνουμε την απεικόνιση του αποτελέσματος των κανονικοποιημένων δεδομένων.

4.2.4: Δημιουργία και εκτέλεση ενός αγωγού

Προσθήκη εκπαιδευτικών ενοτήτων

Είναι κοινή πρακτική να εκπαιδεύεται ένα μοντέλο χρησιμοποιώντας ένα υποσύνολο των δεδομένων, κρατώντας παράλληλα κάποια δεδομένα με τα οποία θα εκπαιδεύσει εκ νέου το μοντέλο. Αυτό μας επιτρέπει να συγκρίνουμε τις ετικέτες που προβλέπει το μοντέλο με τις πραγματικές γνωστές ετικέτες στο αρχικό σύνολο δεδομένων. Επομένως, συνεχίζουμε καθώς θα επεκτείνουμε την αγωγό εκπαίδευσης όπως φαίνεται εδώ:



Ακολουθούμε τα παρακάτω βήματα, χρησιμοποιώντας την παραπάνω εικόνα για αναφορά καθώς προσθέτουμε και ρυθμίζουμε τις απαιτούμενες ενότητες.

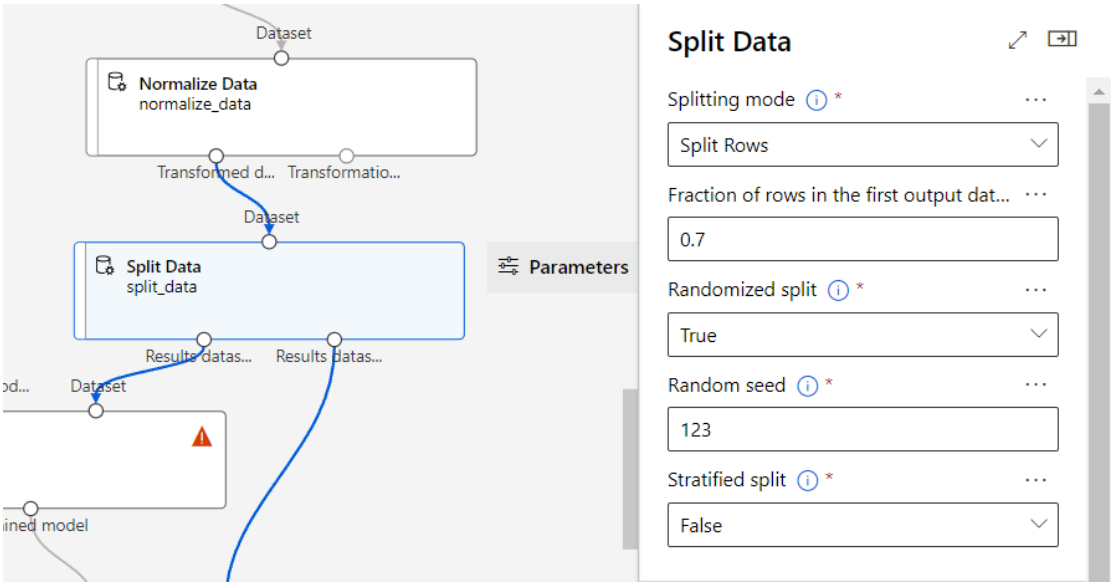
1. Ανοίγουμε τον αγωγό αυτόματης εκτίμησης τιμών που δημιουργήσαμε στην προηγούμενη ενότητα, αν δεν είναι ήδη ανοιχτός.

2. Στο παράθυρο στα αριστερά, στην ενότητα *Data Transformation*, σέρνουμε μια ενότητα *Split Data* (Διαχωρισμός δεδομένων) στον καμβά κάτω από την ενότητα *Normalize Data* (Κανονικοποίηση δεδομένων). Στη συνέχεια, συνδέουμε την έξοδο Transformed Dataset

(αριστερή) της μονάδας *Normalize Data* με την είσοδο της μονάδας *Split Data* (Διαχωρισμός δεδομένων).

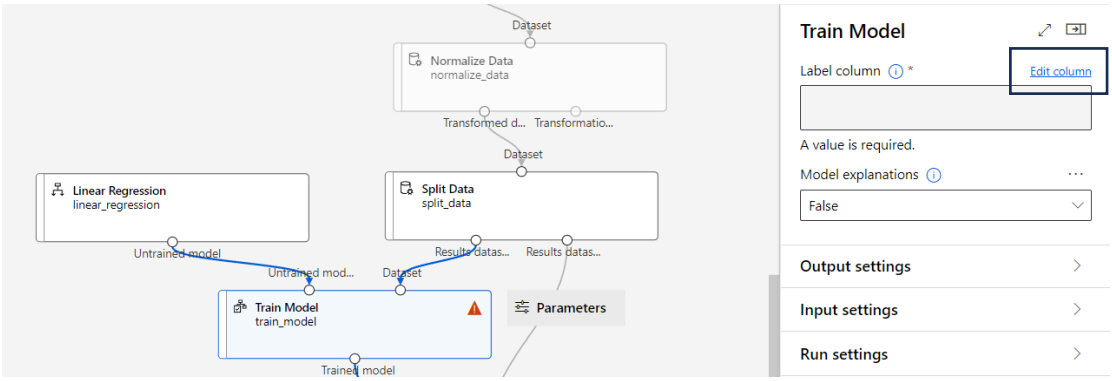
3. Επιλέγουμε την ενότητα *Split Data* με διπλό κλικ και διαμορφώνουμε τις ρυθμίσεις της ως εξής:

- **Splitting mode**: Split Rows
- **Fraction of rows in the first output dataset**: 0.7
- **Random seed**: 123
- **Stratified split**: False

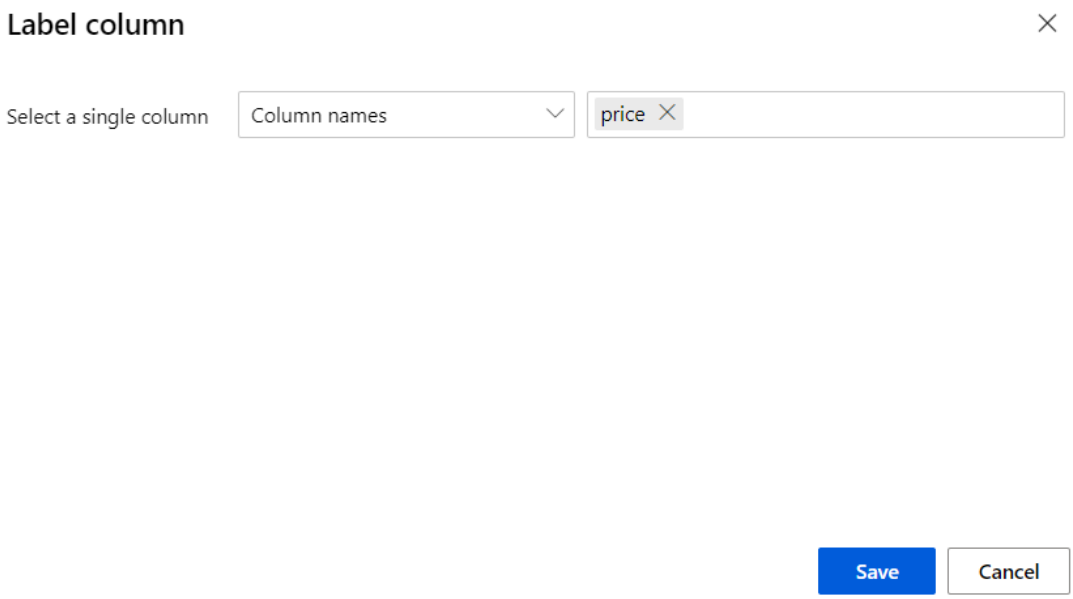


4. Αναπτύσσουμε την ενότητα *Model Training,* στο παράθυρο στα αριστερά και σέρνουμε μια ενότητα *Train Model* στον καμβά, κάτω από την ενότητα *Split Data*. Στη συνέχεια, συνδέουμε την έξοδο Result dataset1 (αριστερά) της μονάδας *Split Data* με την είσοδο *Dataset* (δεξιά) της μονάδας *Train Model*.

5. Το μοντέλο που εκπαιδεύουμε θα προβλέψει την τιμή της Price, οπότε επιλέγουμε την ενότητα *Train Model* με διπλό κλικ και τροποποιούμε τις ρυθμίσεις πατώντας *Edit column*.



Επιλέγουμε Column names και πληκτρολογούμε το όνομα της στήλης που θα προβλέψουμε, δηλαδή price.



6. Η ετικέτα price που θα προβλέψει το μοντέλο είναι μια αριθμητική τιμή, οπότε πρέπει να εκπαιδεύσουμε το μοντέλο χρησιμοποιώντας έναν αλγόριθμο παλινδρόμησης. Αναπτύσσουμε την ενότητα *Machine Learning Algorithms* (Αλγόριθμοι μηχανικής μάθησης) και στην ενότητα *Regression* (Παλινδρόμηση) σέρνουμε την ενότητα *Linear Regression* (Γραμμική παλινδρόμηση) στον καμβά, στα αριστερά της ενότητας *Split Data* και πάνω από την ενότητα *Train Model*. Στη συνέχεια, συνδέουμε την έξοδό της με την είσοδο *Untrained model* (αριστερά) της ενότητας *Train Model*.

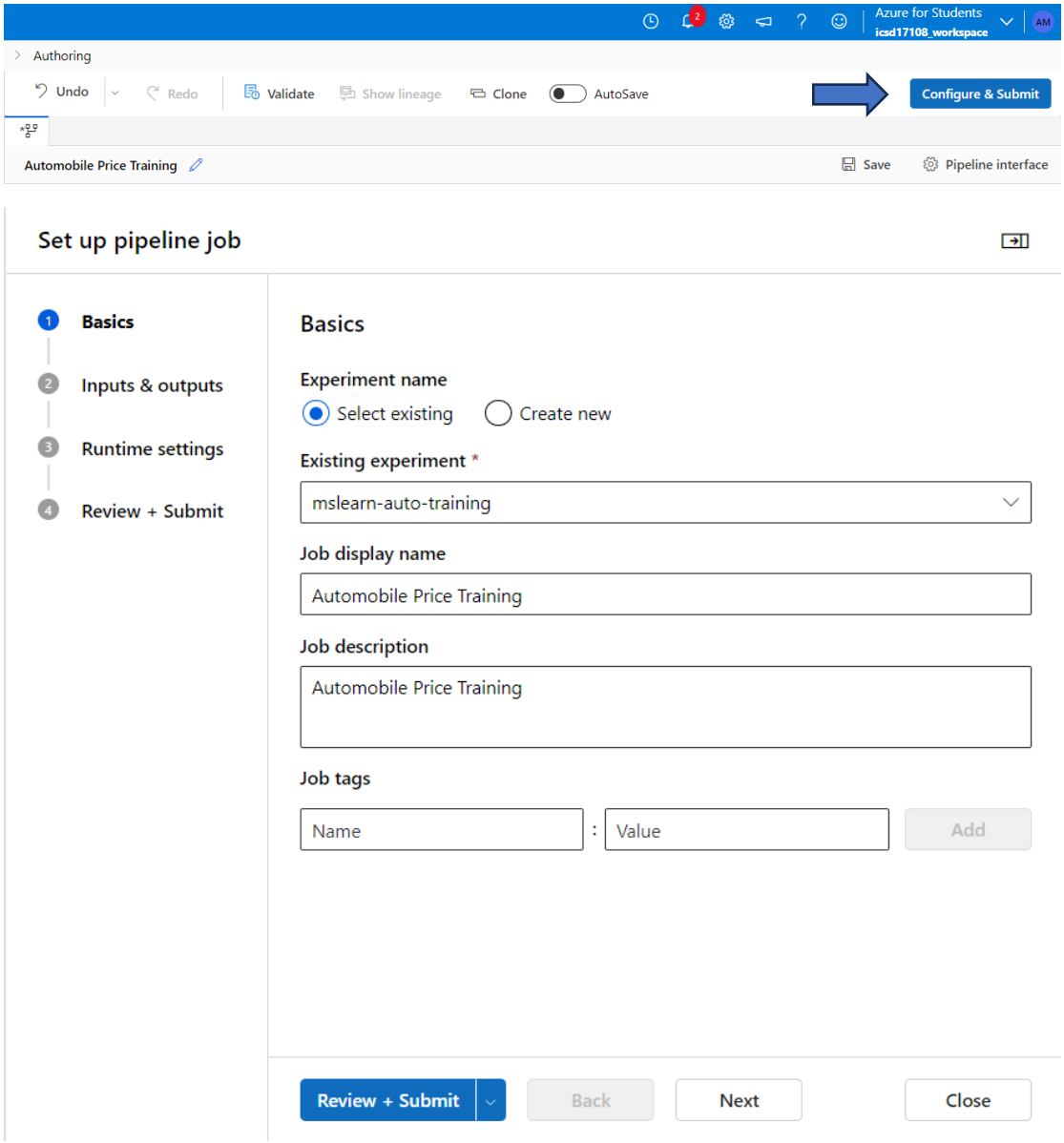
7. Για να δοκιμάσουμε το εκπαιδευμένο μοντέλο, πρέπει να το χρησιμοποιήσουμε για να βαθμολογήσουμε το σύνολο δεδομένων επικύρωσης που κρατήσαμε πίσω όταν χωρίσαμε τα αρχικά δεδομένα - με άλλα λόγια, να προβλέψουμε ετικέτες για τα χαρακτηριστικά στο σύνολο δεδομένων επικύρωσης. Αναπτύσσουμε την ενότητα *Model Scoring & Evaluation* (Βαθμολόγηση και αξιολόγηση μοντέλου) και σέρνουμε την ενότητα *Score Model* (Βαθμολόγηση μοντέλου) στον καμβά, κάτω από την ενότητα *Train Model*. Στη συνέχεια, συνδέουμε την έξοδο της ενότητας *Train Model* στην είσοδο *Trained model* (αριστερά) της ενότητας *Score Model* και σέρνουμε την έξοδο *Results dataset2* (δεξιά) της ενότητας *Split Data* στην είσοδο *Dataset* (δεξιά) της ενότητας *Score Model*.

8. Επικυρώνουμε ότι ο αγωγός μας έχει την τελική μορφή που είδαμε προηγουμένως.

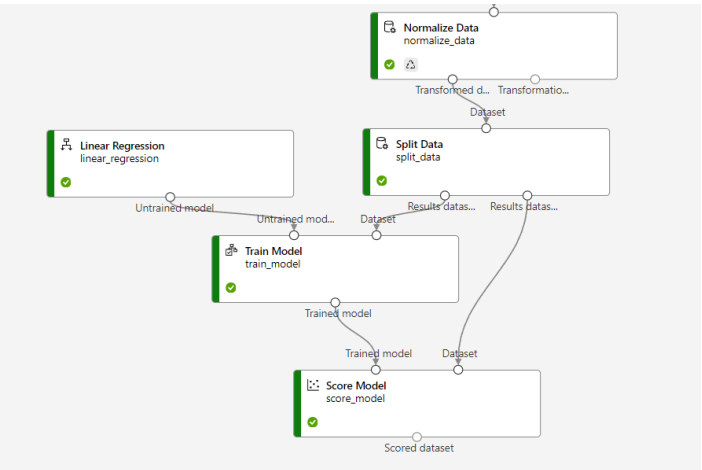
Εκτέλεση του εκπαιδευτικού αγωγού

Τώρα είμαστε έτοιμοι να εκτελέσουμε τον εκπαιδευτικό αγωγό και να εκπαιδεύσουμε το μοντέλο μας, ακολουθώντας τα εξής βήματα:

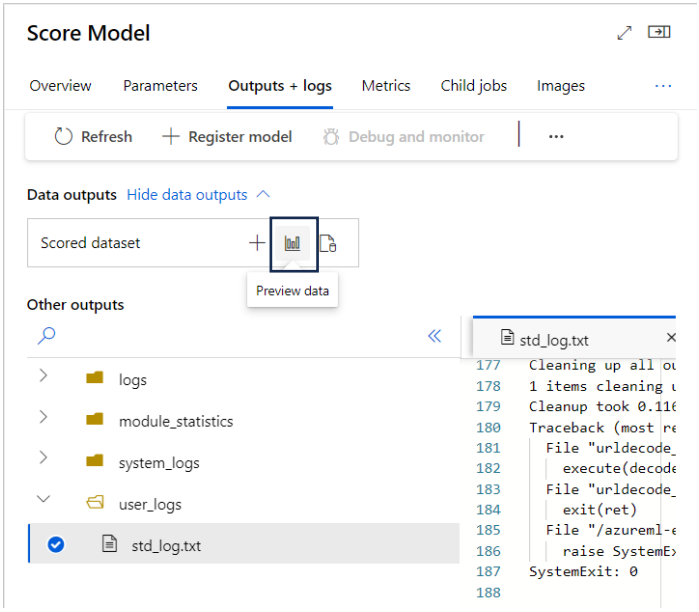
1. Επιλέγουμε *Configure & Submit* και εκτελούμε τον αγωγό χρησιμοποιώντας το υπάρχον πείραμα με όνομα mslearn-auto-training. Στη συνέχεια, πατάμε *Review & Submit* κι έπειτα *Submit*.



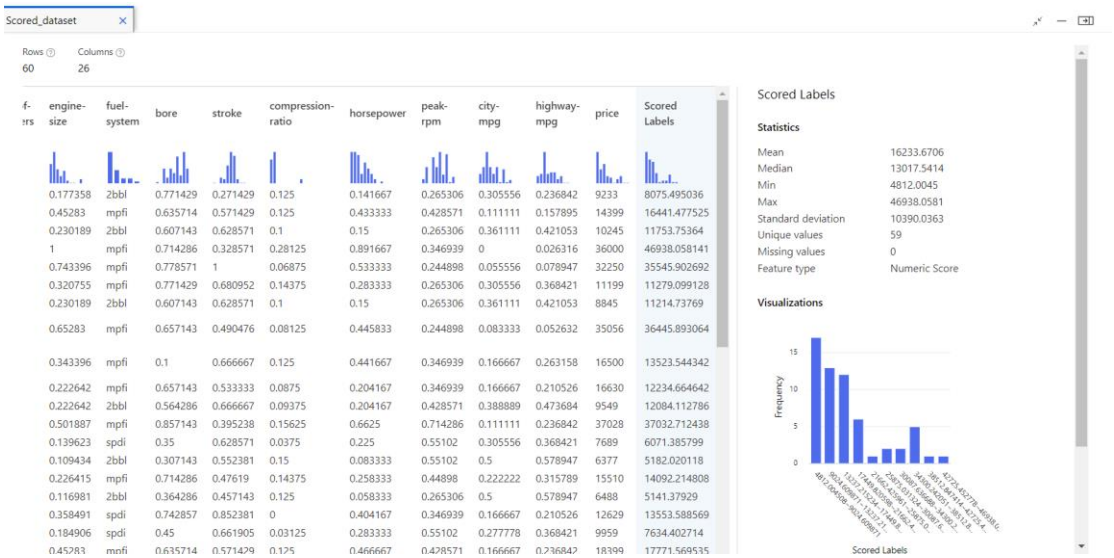
2. Περιμένουμε να ολοκληρωθεί η εκτέλεση του πειράματος, το οποίο μπορεί να διαρκέσει 5 λεπτά ή και περισσότερο.



3. Όταν ολοκληρωθεί η εκτέλεση του πειράματος, επιλέγουμε την ενότητα *Score Model* με διπλό κλικ και στο παράθυρο ρυθμίσεων, στην καρτέλα *Outputs + logs*, στην ενότητα *Data outputs* στην ενότητα *Scored dataset* (Σύνολο δεδομένων με βαθμολογία), επιλέγουμε Show data outputs και Preview data όπως προηγουμένως.



4. Κάνουμε κύλιση προς τα δεξιά και παρατηρούμε ότι δίπλα στη στήλη price, η οποία περιέχει τις γνωστές πραγματικές τιμές της ετικέτας, υπάρχει μια νέα στήλη με όνομα Scored labels, η οποία περιέχει τις προβλεπόμενες τιμές της ετικέτας.



1. Κλείνουμε το παράθυρο απεικόνισης των αποτελεσμάτων του μοντέλου βαθμολογίας. Το μοντέλο προβλέπει τιμές για την ετικέτα τιμής, αλλά πόσο αξιόπιστες είναι οι προβλέψεις του; Για να το εκτιμήσουμε αυτό, πρέπει να αξιολογήσουμε το μοντέλο.

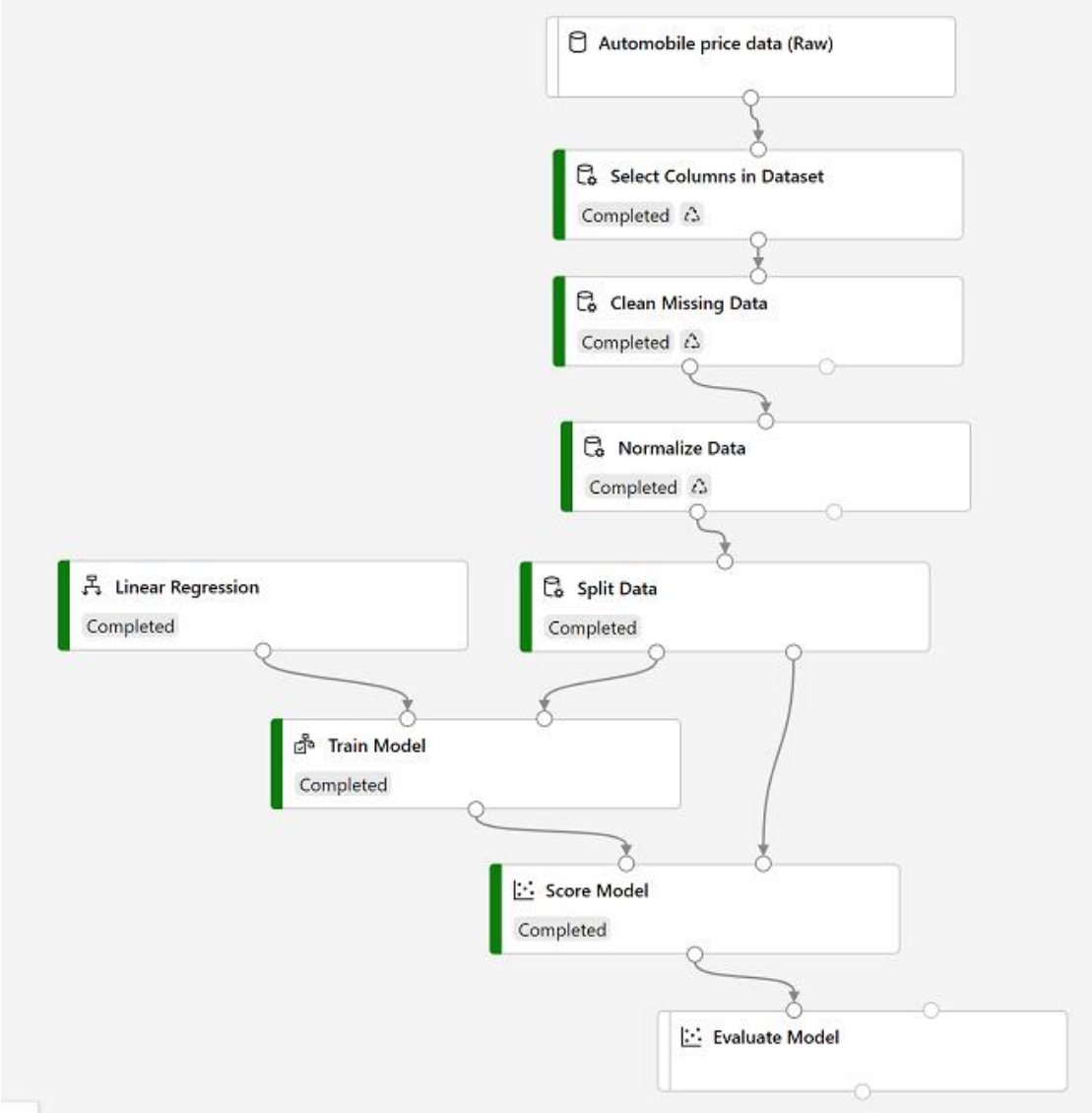
4.2.5 Αξιολόγηση μοντέλου παλινδρόμησης

Για να αξιολογήσουμε ένα μοντέλο παλινδρόμησης, θα μπορούσαμε απλώς να συγκρίνουμε τις προβλεπόμενες ετικέτες με τις πραγματικές ετικέτες στο σύνολο δεδομένων επικύρωσης που κρατήσαμε κατά τη διάρκεια της εκπαίδευσης, αλλά αυτή είναι μια ανακριβής διαδικασία και δεν παρέχει μια απλή μετρική που μπορούμε να χρησιμοποιήσουμε για να συγκρίνουμε την απόδοση πολλαπλών μοντέλων. Επομένως, προσθέτουμε μια ενότητα *Evaluate Model* (Αξιολόγηση μοντέλου) ως εξής:

1. Ανοίγουμε τον αγωγό αυτόματης εκτίμησης τιμών που δημιουργήσαμε στην προηγούμενη ενότητα, αν δεν είναι ήδη ανοιχτός.

2. Στο παράθυρο στα αριστερά, στην ενότητα *Model Scoring & Evaluation*, σέρνουμε την ενότητα *Evaluate Model* στον καμβά, κάτω από την ενότητα *Score Model*, και συνδέουμε την έξοδο της ενότητας *Score Model* με την είσοδο *Scored dataset* (αριστερά) της ενότητας *Evaluate Model*.

3. Βεβαιωνόμαστε ότι ο αγωγός μας είναι κάπως έτσι:



4. Επιλέγουμε *Configure & Submit* και εκτελούμε τον αγωγό χρησιμοποιώντας το υπάρχον πείραμα με όνομα mslearn-auto-training. Στη συνέχεια, πατάμε *Review & Submit* κι έπειτα *Submit*. Περιμένετε να ολοκληρωθεί η εκτέλεση του πειράματος.

6. Όταν ολοκληρωθεί η εκτέλεση του πειράματος, επιλέξτε την ενότητα *Evaluate Model* με διπλό κλικ και στο παράθυρο ρυθμίσεων, στην καρτέλα *Outputs + logs*, στην ενότητα *Data outputs*, στην ενότητα Evaluation results προβάλουμε τα αποτελέσματα. Αυτά περιλαμβάνουν τις ακόλουθες μετρήσεις απόδοσης παλινδρόμησης:

- **Μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE)**: Η μέση διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών. Η τιμή αυτή βασίζεται στις ίδιες μονάδες με την ετικέτα, στην προκειμένη περίπτωση σε δολάρια. Όσο χαμηλότερη είναι αυτή η τιμή, τόσο καλύτερα προβλέπει το μοντέλο.
- **Μέσο τετραγωνικό σφάλμα (Root Mean Squared Error - RMSE)**: Η τετραγωνική ρίζα της μέσης τετραγωνικής διαφοράς μεταξύ προβλεπόμενων και πραγματικών τιμών. Το αποτέλεσμα είναι μια μετρική που βασίζεται στην ίδια μονάδα με την ετικέτα

(δολάρια). Σε σύγκριση με το MAE (παραπάνω), μια μεγαλύτερη διαφορά υποδηλώνει μεγαλύτερη διακύμανση στα επιμέρους σφάλματα (για παράδειγμα, με ορισμένα σφάλματα να είναι πολύ μικρά, ενώ άλλα να είναι μεγάλα).

- **Σχετικό τετραγωνικό σφάλμα (Relative Squared Error - RSE)**: Μια σχετική μετρική μεταξύ 0 και 1 με βάση το τετράγωνο των διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Όσο πιο κοντά στο 0 είναι αυτή η μετρική, τόσο καλύτερη είναι η απόδοση του μοντέλου. Επειδή αυτή η μετρική είναι σχετική, μπορεί να χρησιμοποιηθεί για τη σύγκριση μοντέλων όπου οι ετικέτες είναι σε διαφορετικές μονάδες.
- **Σχετικό απόλυτο σφάλμα (Relative Absolute Error - RAE)**: Μια σχετική μετρική μεταξύ 0 και 1 με βάση τις απόλυτες διαφορές μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Όσο πιο κοντά στο 0 είναι αυτή η μετρική, τόσο καλύτερη είναι η απόδοση του μοντέλου. Όπως το RSE, αυτή η μετρική μπορεί να χρησιμοποιηθεί για τη σύγκριση μοντέλων όπου οι ετικέτες είναι σε διαφορετικές μονάδες.
- **Συντελεστής προσδιορισμού (Coefficient of Determination)**: Αυτή η μετρική αναφέρεται συνηθέστερα ως R-τετράγωνο και συνοψίζει πόσο μεγάλο μέρος της διακύμανσης μεταξύ προβλεπόμενων και πραγματικών τιμών εξηγείται από το μοντέλο. Όσο πιο κοντά στο 1 είναι αυτή η τιμή, τόσο καλύτερη είναι η απόδοση του μοντέλου

7. Κλείνουμε το παράθυρο απεικόνισης αποτελεσμάτων *Evaluate Model*.

4.3 Σενάριο 3ο | Εκπαίδευση ενός μοντέλου ταξινόμησης με τη χρήση Azure Machine Learning

Στο συγκεκριμένο σενάριο θα δούμε πώς μπορούμε να εκπαιδεύσουμε και να αναπτύξουμε ένα μοντέλο ταξινόμησης με το Microsoft Azure Machine Learning. Η ταξινόμηση είναι μια μορφή μηχανικής μάθησης που χρησιμοποιείται για να προβλέψει σε ποια κατηγορία ή τάξη ανήκει ένα στοιχείο. Για παράδειγμα, μια κλινική υγείας μπορεί να χρησιμοποιήσει τα

χαρακτηριστικά ενός ασθενούς, όπως η ηλικία, το βάρος, η αρτηριακή πίεση κλπ., για να προβλέψει εάν ο ασθενής κινδυνεύει από διαβήτη. Στην περίπτωση αυτή, τα χαρακτηριστικά του ασθενούς είναι τα χαρακτηριστικά και η ετικέτα είναι μια ταξινόμηση είτε μηδέν είτε μία που αντιπροσωπεύει εάν ο ασθενής είναι μη διαβητικός ή διαβητικός. Η ταξινόμηση είναι ένα παράδειγμα εποπτευόμενης τεχνικής μηχανικής μάθησης στην οποία εκπαιδεύεται ένα μοντέλο χρησιμοποιώντας δεδομένα που περιλαμβάνουν τόσο τα χαρακτηριστικά όσο και τις γνωστές τιμές για την ετικέτα. Με την ταξινόμηση, το μοντέλο μαθαίνει να προσαρμόζει τους συνδυασμούς χαρακτηριστικών στην ετικέτα. Στο παράδειγμα της κλινικής υγείας μας, συλλέγονται μετρήσεις όπως το βάρος, το ύψος, το επίπεδο γλυκόζης, η αρτηριακή πίεση κλπ. για κάθε ασθενή. Γνωρίζουμε επίσης εάν ο ασθενής που μετράμε είναι διαβητικός ή όχι. Σε αυτή την περίπτωση, οι τάξεις είναι εάν ο ασθενής είναι διαβητικός ή μη διαβητικός. Όταν ολοκληρωθεί η εκπαίδευση, μπορούμε να χρησιμοποιήσουμε το εκπαιδευμένο μοντέλο για να προβλέψουμε ετικέτες για νέα αντικείμενα για τα οποία η ετικέτα είναι άγνωστη. Όταν λαμβάνονται μετρήσεις ενός νέου ασθενούς, το εκπαιδευμένο μοντέλο χρησιμοποιεί αυτά τα χαρακτηριστικά για να προβλέψει την πιθανότητα ο ασθενής να είναι διαβητικός ή μη διαβητικός.

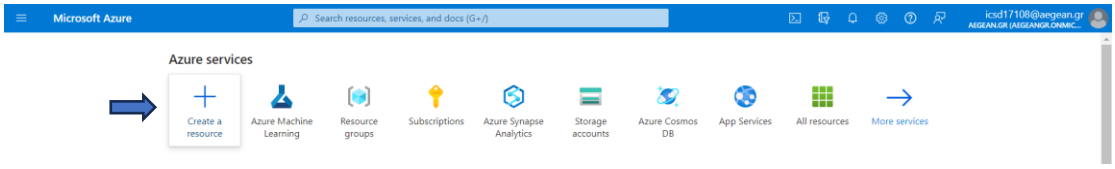
4.3.1 Δημιουργία ενός χώρου εργασίας

Προκειμένου να χρησιμοποιήσουμε το Azure ML, δημιουργούμε έναν χώρο εργασίας (workspace) στη συνδρομή μας στην πλατφόρμα. Στη συνέχεια, μπορούμε να χρησιμοποιήσουμε αυτόν τον χώρο εργασίας για τη διαχείριση δεδομένων, υπολογιστικών πόρων, κώδικα, μοντέλων και άλλων αντικειμένων που σχετίζονται με τους φόρτους εργασίας μηχανικής μάθησης.

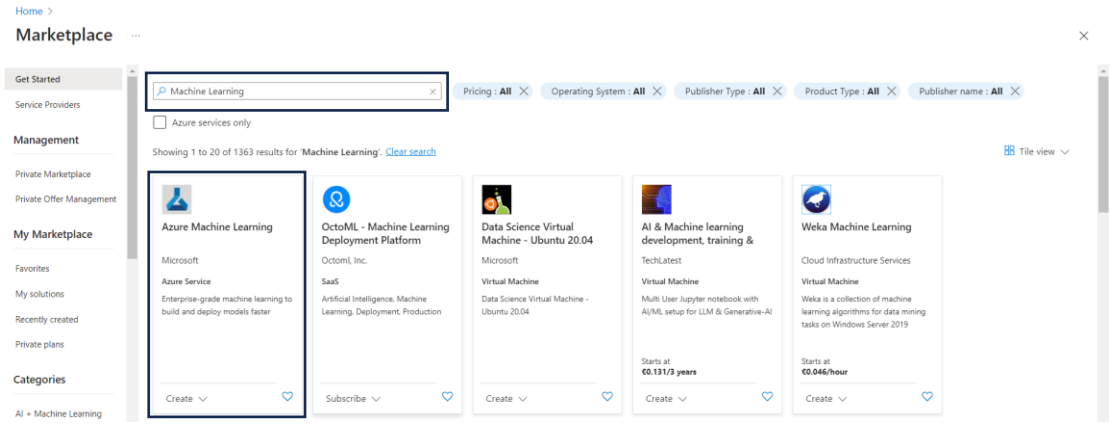
Σε περίπτωση που δεν διαθέτουμε ήδη ένα workspace, ακολουθούμε τα παρακάτω βήματα για να δημιουργήσουμε ένα χώρο εργασίας:

1. Συνδεόμαστε στην πύλη Microsoft Azure χρησιμοποιώντας τα διαπιστευτήριά μας.

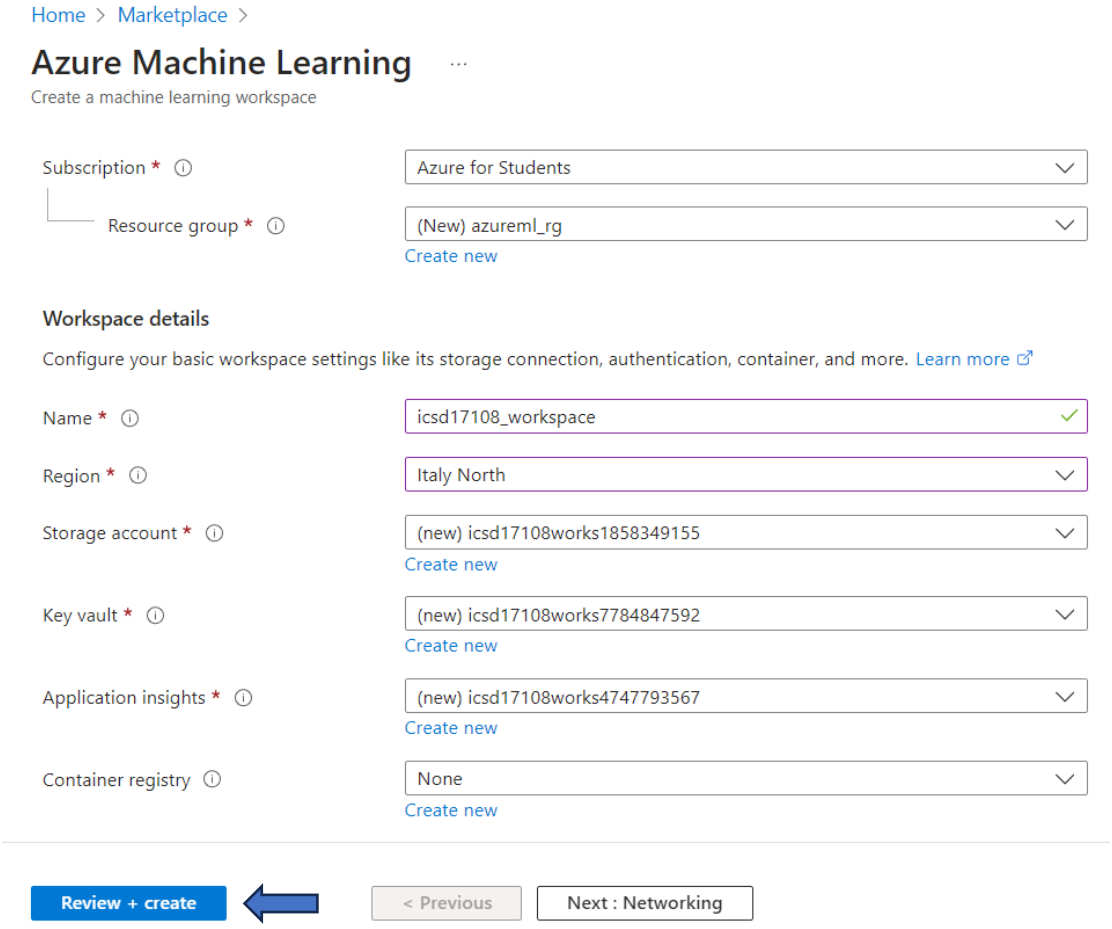
2. Επιλέγουμε + *Create a resource* (Δημιουργία πόρου).



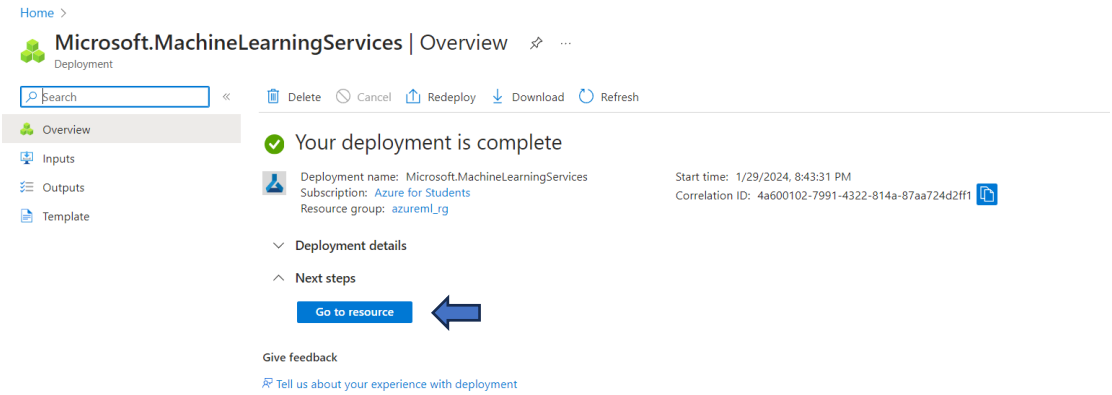
Έπειτα κάνουμε αναζήτηση *Machine Learning* και δημιουργούμε έναν νέο πόρο *Azure Machine Learning* με τις ακόλουθες ρυθμίσεις:



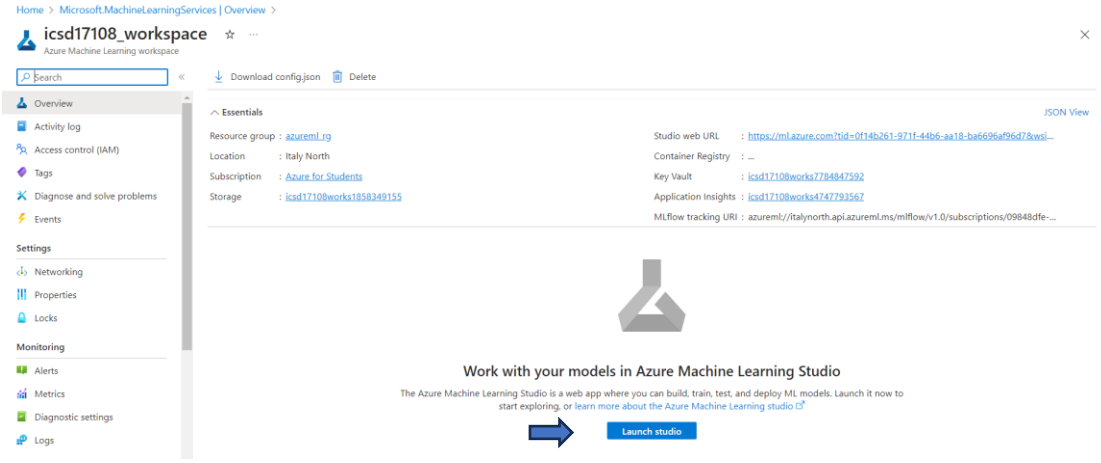
- **Subscription**: Η συνδρομή μας στο Azure.
- **Resource group**: Δημιουργούμε νέο ή επιλέγουμε μια ομάδα πόρων.
- **Name**: Εισάγουμε ένα μοναδικό όνομα για τον χώρο εργασίας μας.
- **Region**: Επιλέγουμε τη γεωγραφική περιοχή που βρίσκεται πιο κοντά σε εμάς.
- **Storage account**, **Key vault**, **Applications insights**: Συμπληρώνονται αυτόματα εφόσον ορίσουμε προηγουμένως το όνομα.



3. Εφόσον πατήσαμε *Review + create* και μετά ξανά *Create*, περιμένουμε μερικά λεπτά έως ότου δημιουργηθεί ο χώρος εργασίας μας. Θα ενημερωθούμε με σχετικό μήνυμα κατά την ολοκλήρωση της διαδικασίας.



4. Στη συνέχεια πατάμε *Go to resource* όπου μεταφερόμαστε στη σελίδα επισκόπησης (Overview) του χώρου εργασίας μας. Τώρα μπορούμε να εκκινήσουμε το Microsoft Azure Machine Learning Studio επιλέγοντας *Launch studio*.

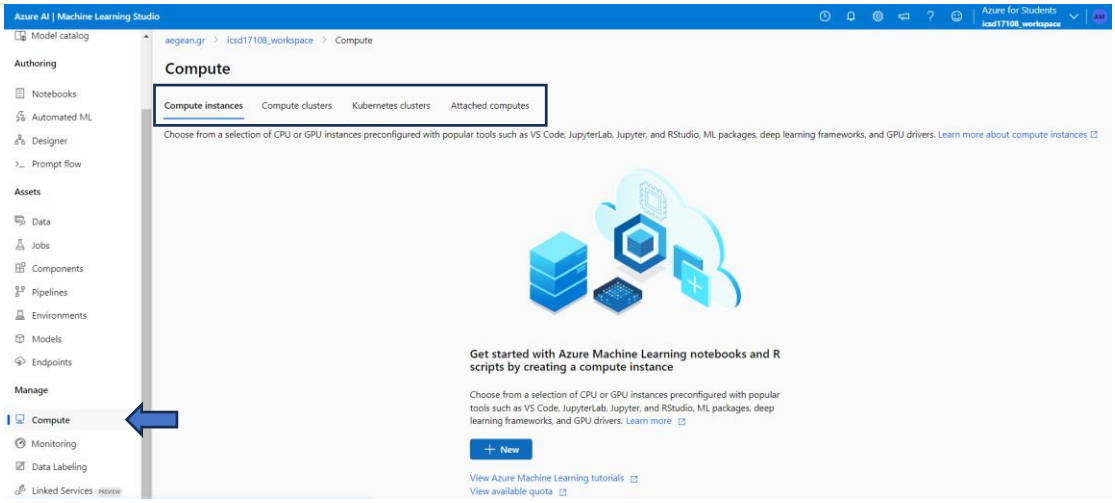


Εναλλακτικά ανοίγουμε μια νέα καρτέλα του προγράμματος περιήγησης και πληκτρολογούμε τη διεύθυνση *https://ml.azure.com*, όπου συνδεόμαστε στο Microsoft Azure Machine Learning Studio χρησιμοποιώντας το λογαριασμό μας στη Microsoft.

4.3.2 Δημιουργία υπολογιστικών πόρων

Οι υπολογιστικοί πόροι που θα χρησιμοποιήσουμε βασίζονται στο cloud, πάνω στους οποίους μπορούμε να εκτελέσουμε τις διαδικασίες εκπαίδευσης μοντέλων και διερεύνησης δεδομένων. Ακολουθούμε τα παρακάτω βήματα για την δημιουργία υπολογιστικών πόρων:

1. Στην πλατφόρμα επιλέγοντας το εικονίδιο ☰ στο επάνω αριστερό μέρος προβάλλονται οι διάφορες σελίδες του περιβάλλοντος εργασίας. Έπειτα, επιλέγουμε τη σελίδα *Compute* που βρίσκεται στην ομάδα *Manage* του μενού.

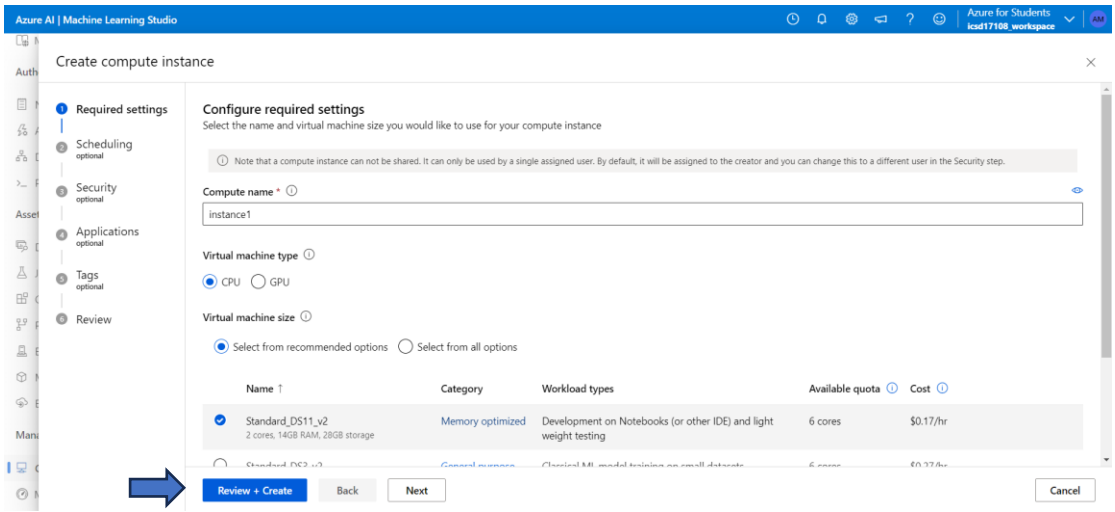


Στη συγκεκριμένη σελίδα διαχειριζόμαστε τους στόχους υπολογισμού για όλες τις δραστηριότητες ανάλυσης και διερεύνησης των δεδομένων μας. Υπάρχουν τέσσερα είδη υπολογιστικών πόρων που μπορούμε να δημιουργήσουμε:

- **Compute Instances (Μονάδες υπολογισμού)**: Οι σταθμοί εργασίας ή ανάπτυξης που μπορούν να χρησιμοποιούν οι επιστήμονες δεδομένων για να εργάζονται με δεδομένα και μοντέλα.
- **Compute Clusters (Συστάδες υπολογισμών)**: Οι επεκτάσιμες συστάδες εικονικών μηχανών για την κατά παραγγελία επεξεργασία πειραματικού κώδικα.
- **Inference Clusters (Συστάδες συμπερασμάτων)**: Οι στόχοι ανάπτυξης για τις υπηρεσίες πρόβλεψης που χρησιμοποιούν τα εκπαιδευμένα μοντέλα.
- **Attached Compute (Επισυναπτόμενος υπολογισμός)**: Οι συνδέσεις με υπάρχοντες υπολογιστικούς πόρους του Microsoft Azure, όπως εικονικές μηχανές ή Azure Databricks Clusters.

2. Στην καρτέλα *Compute Instances*, θα προσθέσουμε μια νέα μονάδα υπολογισμού, πατώντας *New*, με τις ακόλουθες ρυθμίσεις:

- **Compute name**: Εισάγουμε ένα μοναδικό όνομα.
- **Virtual machine type**: CPU
- **Virtual Machine size**: Standard_DS11_v2. Σε περίπτωση που δεν το βρίσκουμε επιλέγουμε *Select from all options* για να αναζητήσουμε και να επιλέξουμε αυτό το μέγεθος μηχανής.

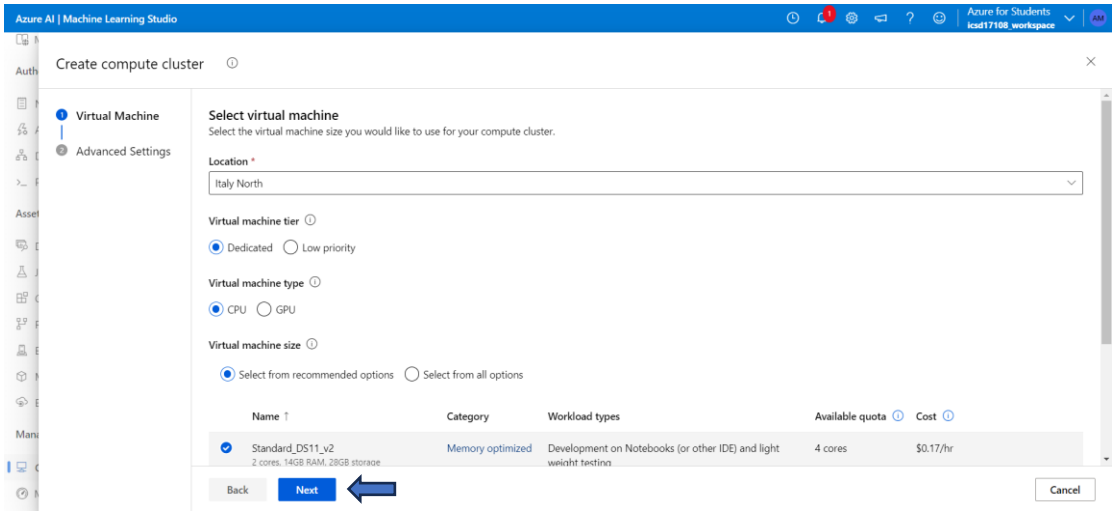


3. Πατάμε Review + Create και μετά ξανά Create, αφού ελέγξουμε την ρύθμιση *Enable SSH Access: Unselected* στην καρτέλα *Security* ή στο τελικό review που θα αναγράφει *no*. Σημειώνεται πως την μονάδα υπολογισμού θα την χρησιμοποιήσουμε ως σταθμό εργασίας από τον οποίο θα δοκιμάσουμε και θα τεστάρουμε το μοντέλο μας.

4. Ενώ δημιουργείται η υπολογιστική μονάδα, πηγαίνουμε στην καρτέλα *Compute Clusters* και προσθέτουμε μια νέα συστάδα υπολογισμών με τις ακόλουθες ρυθμίσεις:

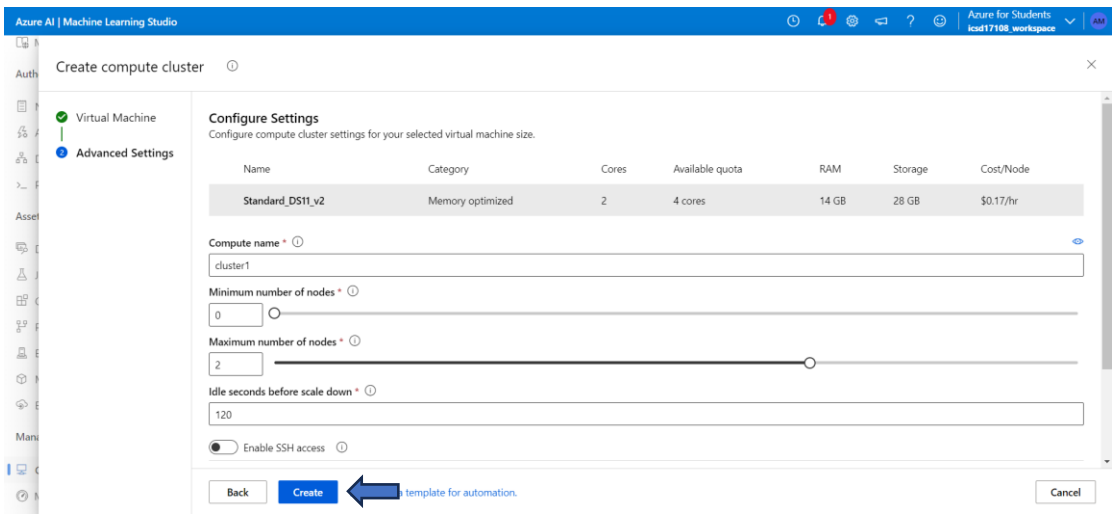
- **Virtual Machine tier:** Dedicated
- **Virtual Machine type:** CPU

- **Virtual Machine size**: Standard_DS11_v2. Σε περίπτωση που δεν το βρίσκουμε επιλέγουμε *Select from all options* για να αναζητήσουμε και να επιλέξουμε αυτό το μέγεθος μηχανής.



Στη συνέχεια πατάμε Next και πηγαίνουμε στην καρτέλα Advanced Settings.

- **Compute name**: Εισάγουμε ένα μοναδικό όνομα.
- **Minimum number of nodes**: 0
- **Maximum number of nodes**: 2
- **Idle seconds before scale down**: 120
- **Enable SSH access**: Unselected.



Τέλος, πατάμε *Create*. Σημειώνεται πως την συγκεκριμένη υπολογιστική μονάδα θα τη χρησιμοποιήσουμε για να εκπαιδεύσουμε ένα μοντέλο μηχανικής μάθησης.

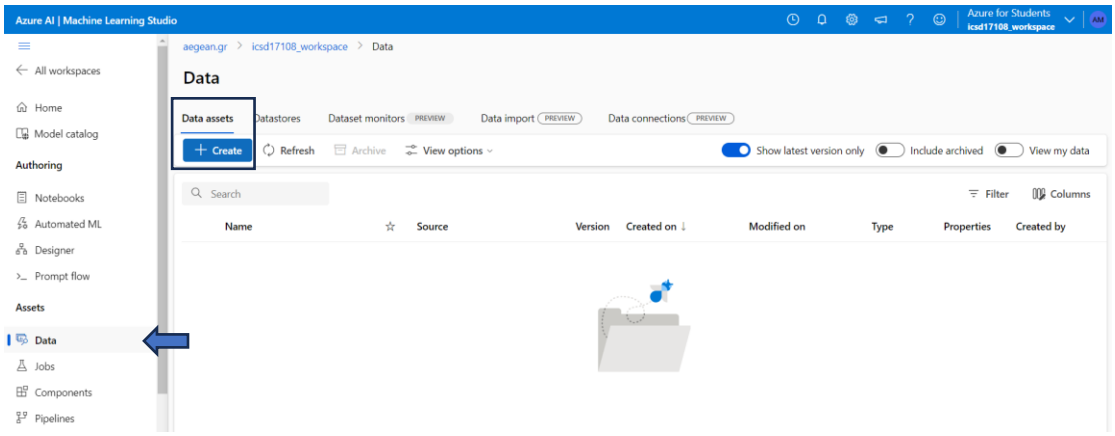
4.5.3 Εξερεύνηση δεδομένων

Για να εκπαιδεύσουμε ένα μοντέλο ταξινόμησης, χρειαζόμαστε ένα σύνολο δεδομένων που περιλαμβάνει ιστορικά χαρακτηριστικά (χαρακτηριστικά της οντότητας για την οποία θέλουμε να κάνουμε πρόβλεψη) και γνωστές τιμές ετικέτας (ο δείκτης κλάσης που θέλουμε να εκπαιδεύσουμε ένα μοντέλο για να προβλέψει).

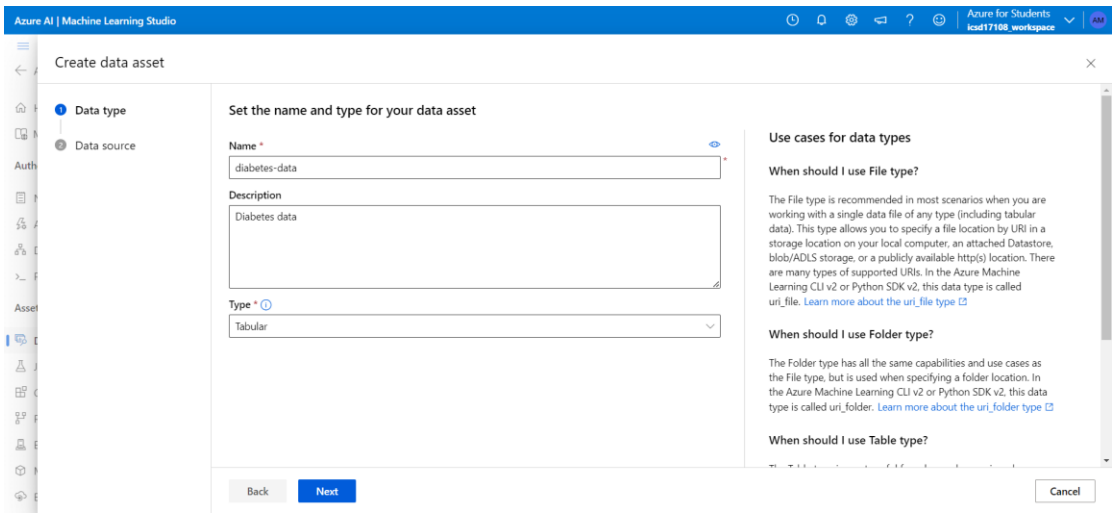
Δημιουργία Dataset

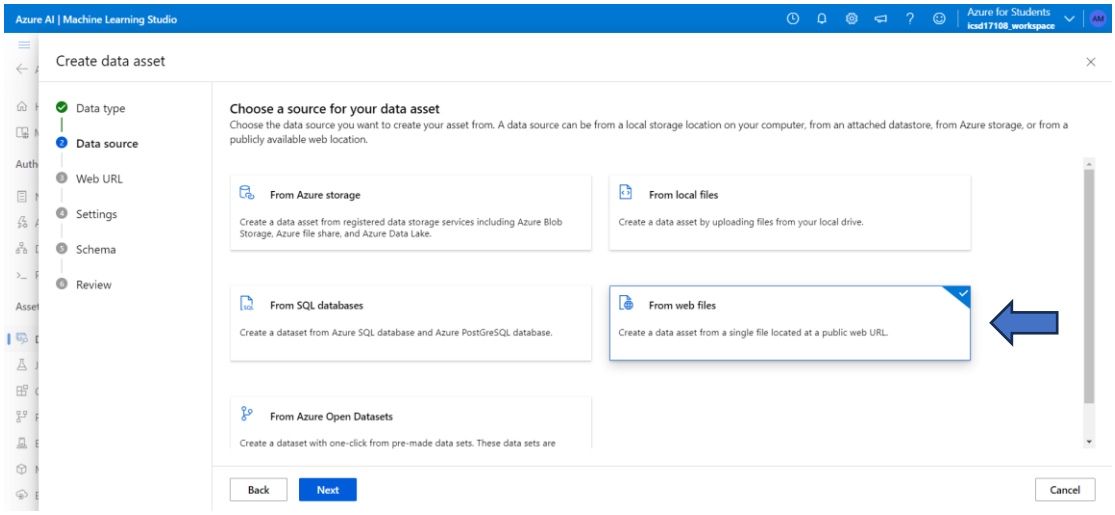
Στο Microsoft Azure ML, τα δεδομένα για την εκπαίδευση του μοντέλου και για άλλες λειτουργίες ενσωματώνονται συνήθως σε ένα αντικείμενο που ονομάζεται dataset.

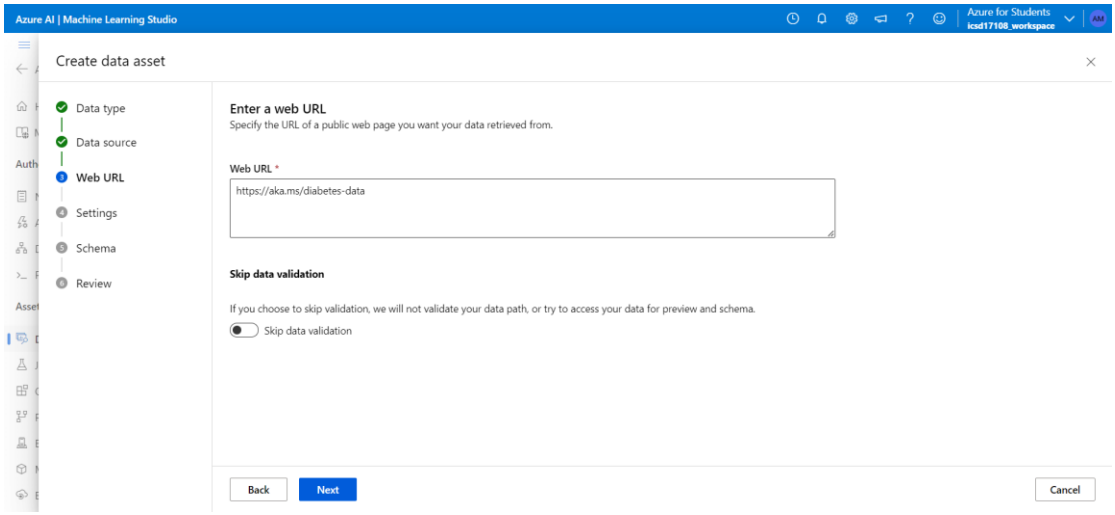
1. Στο Azure Machine Learning studio, προβάλουμε τη σελίδα Data. Τα datasets ή data assets αντιπροσωπεύουν συγκεκριμένα αρχεία δεδομένων ή πίνακες που σκοπεύουμε να χρησιμοποιήσουμε στο Azure ML.

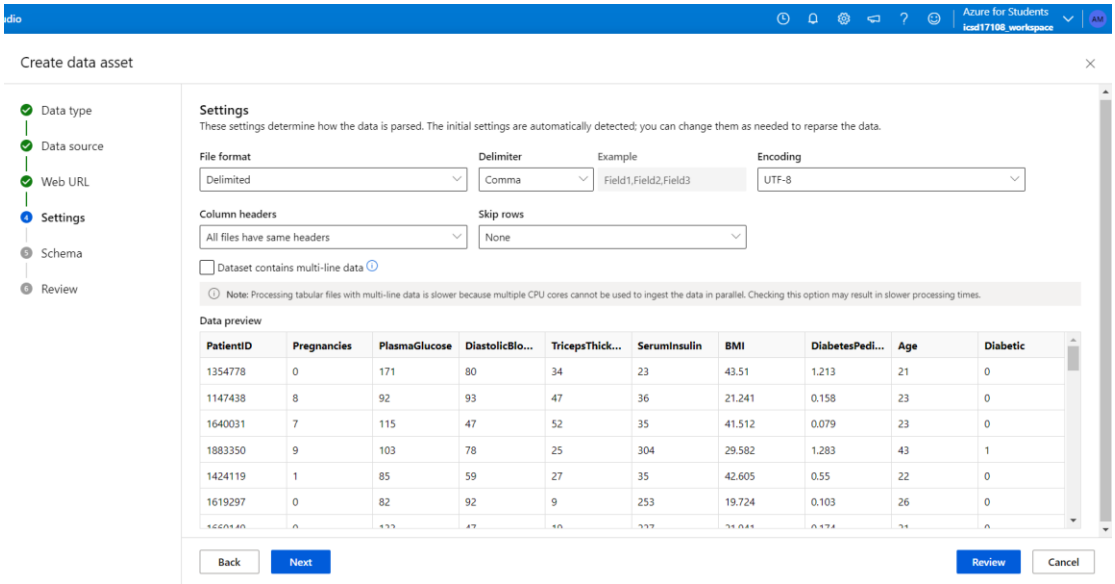


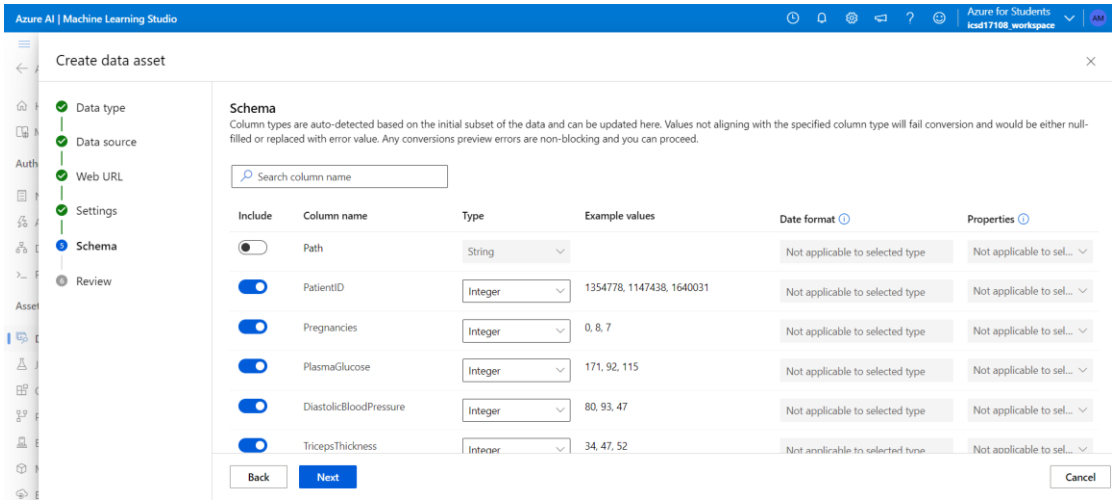
2. Δημιουργούμε ένα dataset από αρχεία στο διαδίκτυο, χρησιμοποιώντας τις ακόλουθες ρυθμίσεις:

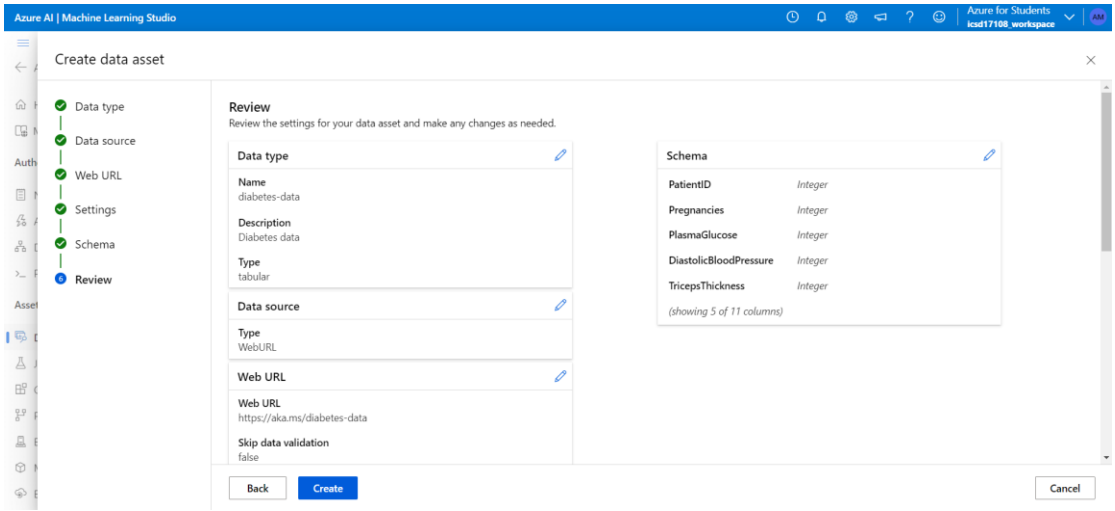



Azure AI | Machine Learning Studio
Azure for Students
icsd17108_workspace
Create data asset
Data type
Data source
Web URL
Settings
Schema
Review
Choose a source for your data asset
Choose the data source you want to create your asset from. A data source can be from a local storage location on your computer, from an attached datastore, from Azure storage, or from a publicly available web location.
From Azure storage
Create a data asset from registered data storage services including Azure Blob Storage, Azure file share, and Azure Data Lake.
From local files
Create a data asset by uploading files from your local drive.
From SQL databases
Create a dataset from Azure SQL database and Azure PostGreSQL database.
From web files
Create a data asset from a single file located at a public web URL.
From Azure Open Datasets
Create a dataset with one-click from pre-made data sets. These data sets are
Back
Next
Cancel

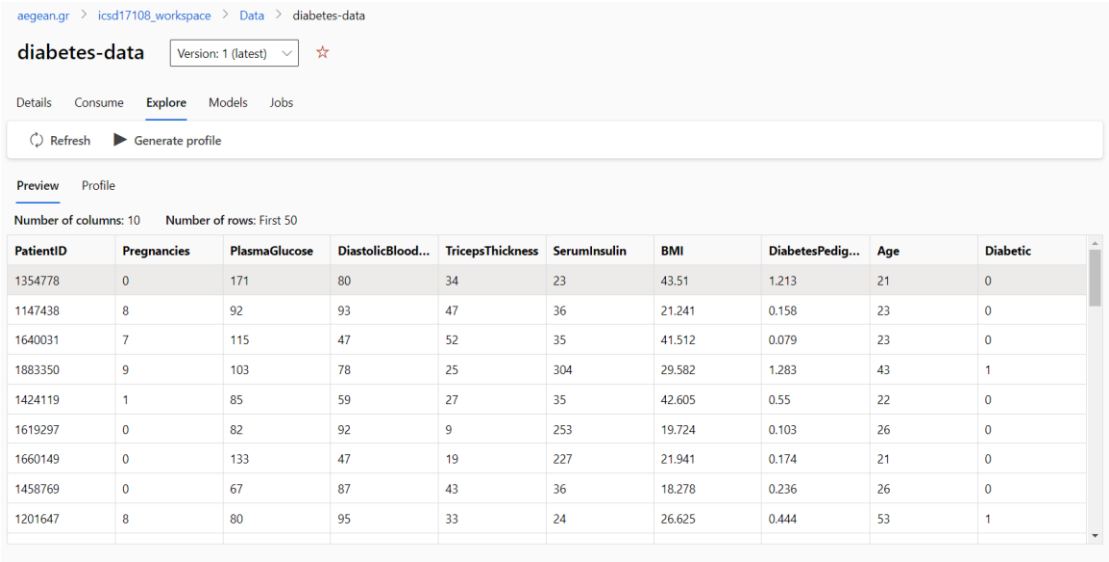

Azure AI | Machine Learning Studio
Azure for Students
icsd17108_workspace
Create data asset
Data type
Data source
Web URL
Settings
Schema
Review
Enter a web URL
Specify the URL of a public web page you want your data retrieved from.
Web URL *
https://aka.ms/diabetes-data
Skip data validation
If you choose to skip validation, we will not validate your data path, or try to access your data for preview and schema.
Skip data validation
Back
Next
Cancel


Azure for Students
icsd17108_workspace
Create data asset
Data type
Data source
Web URL
Settings
Schema
Review
Settings
These settings determine how the data is parsed. The initial settings are automatically detected; you can change them as needed to reparse the data.
File format
Delimited
Delimiter
Comma
Example
Field1,Field2,Field3
Encoding
UTF-8
Column headers
All files have same headers
Skip rows
None
Dataset contains multi-line data
Note: Processing tabular files with multi-line data is slower because multiple CPU cores cannot be used to ingest the data in parallel. Checking this option may result in slower processing times.
Data preview
PatientID Pregnancies PlasmaGlucose DiastolicBlo... TricepsThick... SerumInsulin BMI DiabetesPedi... Age Diabetic
1354778 0 171 80 34 23 43.51 1.213 21 0
1147438 8 92 93 47 36 21.241 0.158 23 0
1640031 7 115 47 52 35 41.512 0.079 23 0
1883350 9 103 78 25 304 29.582 1.283 43 1
1424119 1 85 59 27 35 42.605 0.55 22 0
1619297 0 82 92 9 253 19.724 0.103 26 0
Back
Next
Review
Cancel





3. Αφού δημιουργηθεί το σύνολο δεδομένων, το ανοίγουμε και πηγαίνουμε στη σελίδα *Explore* για να δούμε ένα δείγμα των δεδομένων. Τα δεδομένα αυτά αντιπροσωπεύουν λεπτομέρειες από ασθενείς που έχουν εξεταστεί για διαβήτη.

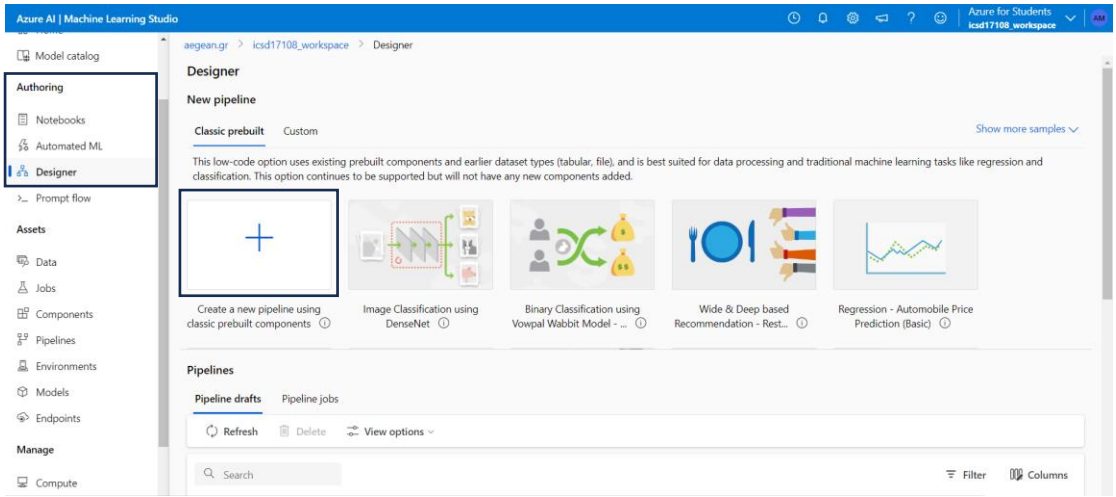


PatientID	Pregnancies	PlasmaGlucose	DiastolicBlood...	TricepsThickness	SerumInsulin	BMI	DiabetesPedig...	Age	Diabetic
1354778	0	171	80	34	23	43.51	1.213	21	0
1147438	8	92	93	47	36	21.241	0.158	23	0
1640031	7	115	47	52	35	41.512	0.079	23	0
1883350	9	103	78	25	304	29.582	1.283	43	1
1424119	1	85	59	27	35	42.605	0.55	22	0
1619297	0	82	92	9	253	19.724	0.103	26	0
1660149	0	133	47	19	227	21.941	0.174	21	0
1458769	0	67	87	43	36	18.278	0.236	26	0
1201647	8	80	95	33	24	26.625	0.444	53	1

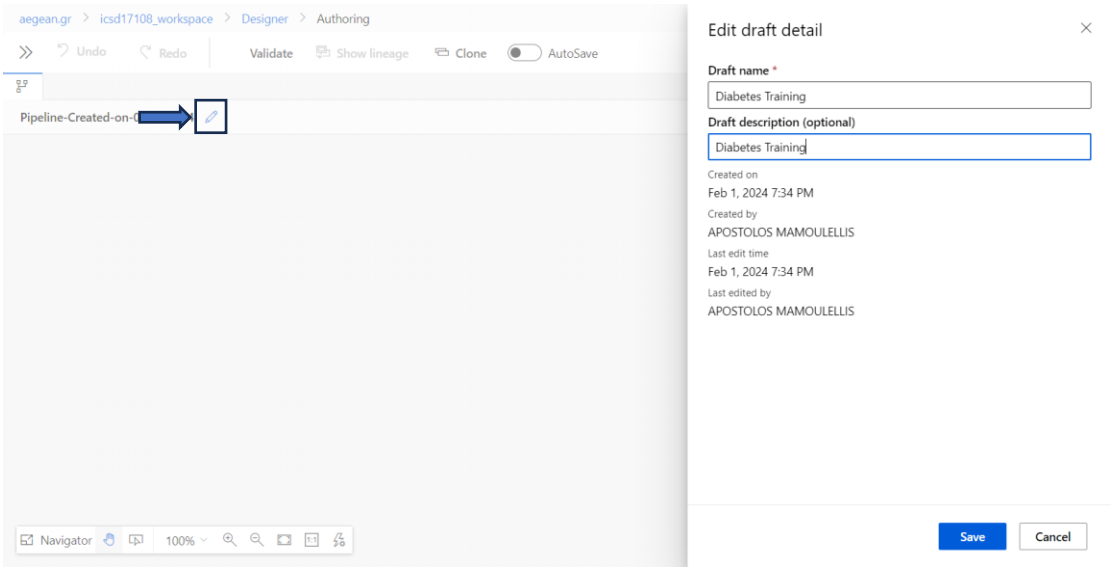
Δημιουργία αγωγού (pipeline)

Αρχικά, για να χρησιμοποιήσουμε τον σχεδιαστή του Microsoft Azure ML, δημιουργούμε έναν αγωγό για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Αυτός ο αγωγός ξεκινά με το σύνολο δεδομένων από το οποίο θέλουμε να εκπαιδεύσουμε το μοντέλο.

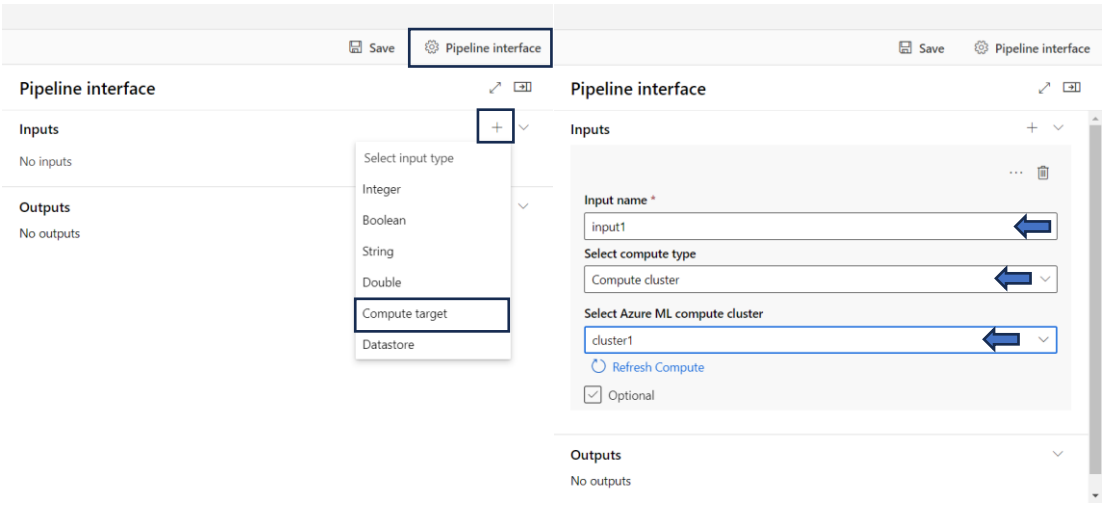
1. Στο μενού βλέπουμε τη σελίδα *Designer* στην ομάδα *Authoring* και επιλέγουμε το εικονίδιο + για να δημιουργήσουμε έναν νέο αγωγό.



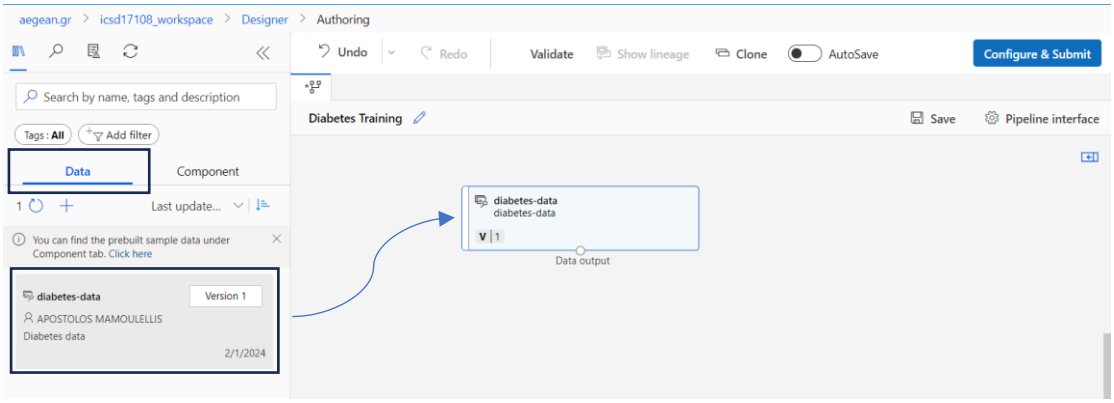
2. Στο παράθυρο ρυθμίσεων, αλλάζουμε το προεπιλεγμένο όνομα του αγωγού (Pipeline-Created-on-date) σε Diabetes Training. Σε περίπτωση που το παράθυρο ρυθμίσεων δεν είναι ορατό, κάνουμε κλικ στο εικονίδιο ✎ δίπλα στο όνομα του αγωγού στο επάνω μέρος.



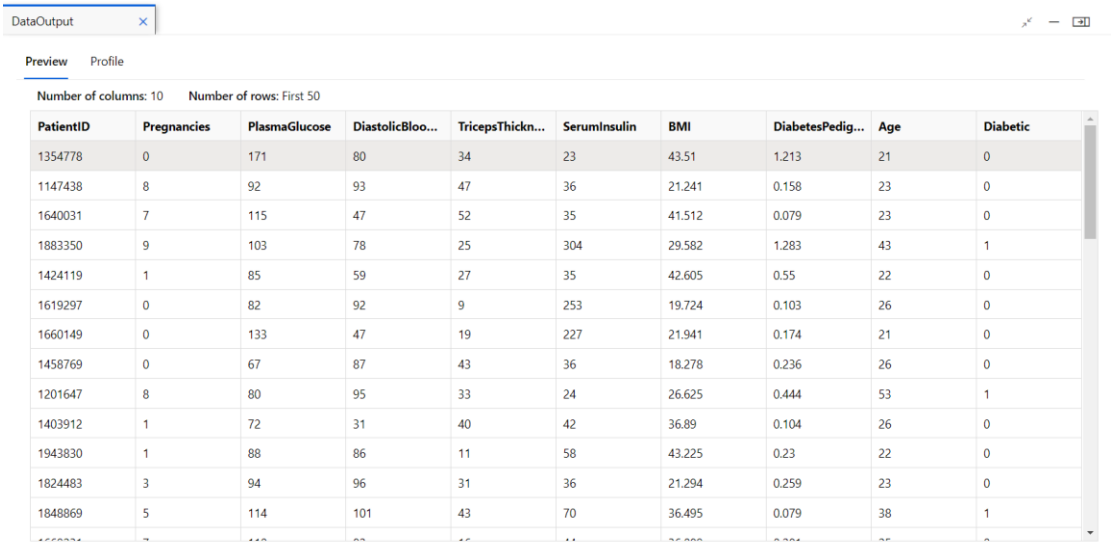
3. Στη συνέχεια πρέπει να καθορίσουμε έναν υπολογιστικό στόχο στον οποίο θα εκτελεστεί ο αγωγός. Στο παράθυρο Pipeline Interface, κάνουμε κλικ στην επιλογή *Select compute target* και επιλέγουμε τη συστάδα υπολογισμών cluster1 που δημιουργήσατε προηγουμένως.



4. Στην αριστερή πλευρά του σχεδιαστή, αναπτύσσουμε την ενότητα Data και σέρνουμε το σύνολο δεδομένων diabetes-data που δημιουργήσαμε προηγουμένως στον καμβά.

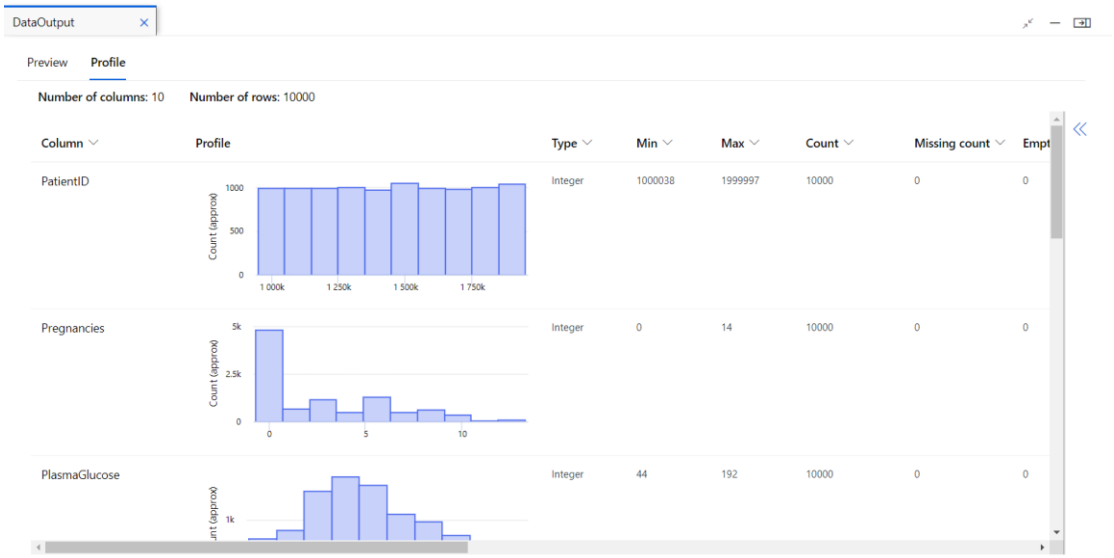


5. Κάνουμε δεξί κλικ στο σύνολο δεδομένων diabetes-data στον καμβά και στο μενού Preview Data, όπου θα μας ανοίξει το παράθυρο *DataOutput*. Στην καρτέλα *Preview* βλέπουμε αναλυτικά τα δεδομένα μας.

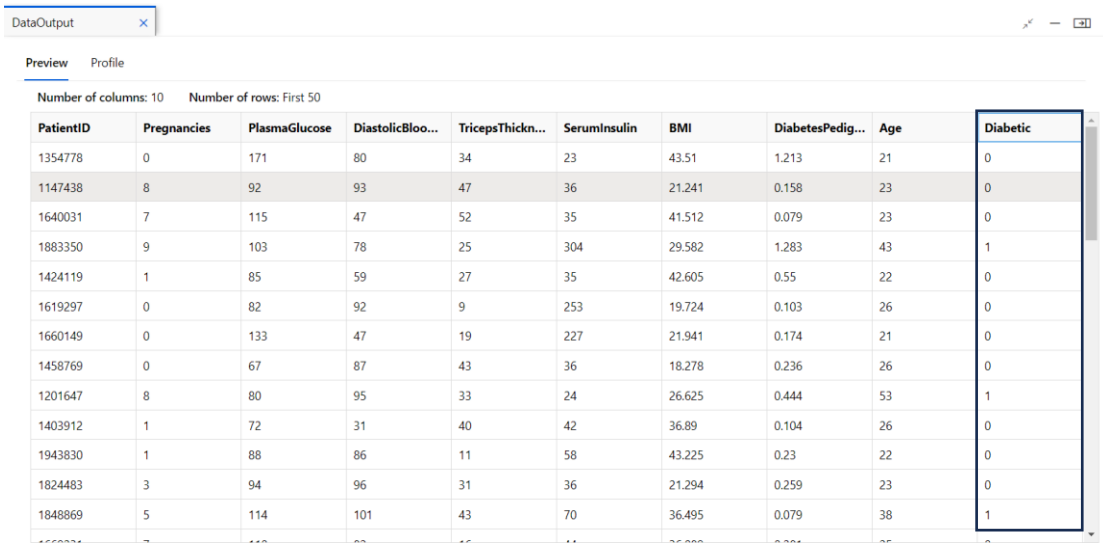


PatientID	Pregnancies	PlasmaGlucose	DiastolicBloo...	TricepsThickn...	SerumInsulin	BMI	DiabetesPedig...	Age	Diabetic
1354778	0	171	80	34	23	43.51	1.213	21	0
1147438	8	92	93	47	36	21.241	0.158	23	0
1640031	7	115	47	52	35	41.512	0.079	23	0
1883350	9	103	78	25	304	29.582	1.283	43	1
1424119	1	85	59	27	35	42.605	0.55	22	0
1619297	0	82	92	9	253	19.724	0.103	26	0
1660149	0	133	47	19	227	21.941	0.174	21	0
1458769	0	67	87	43	36	18.278	0.236	26	0
1201647	8	80	95	33	24	26.625	0.444	53	1
1403912	1	72	31	40	42	36.89	0.104	26	0
1943830	1	88	86	11	58	43.225	0.23	22	0
1824483	3	94	96	31	36	21.294	0.259	23	0
1848869	5	114	101	43	70	36.495	0.079	38	1

6. Επίσης στην καρτέλα Profile, μπορούμε να δούμε τις κατανομές των διαφόρων στηλών ως ιστογράμματα.

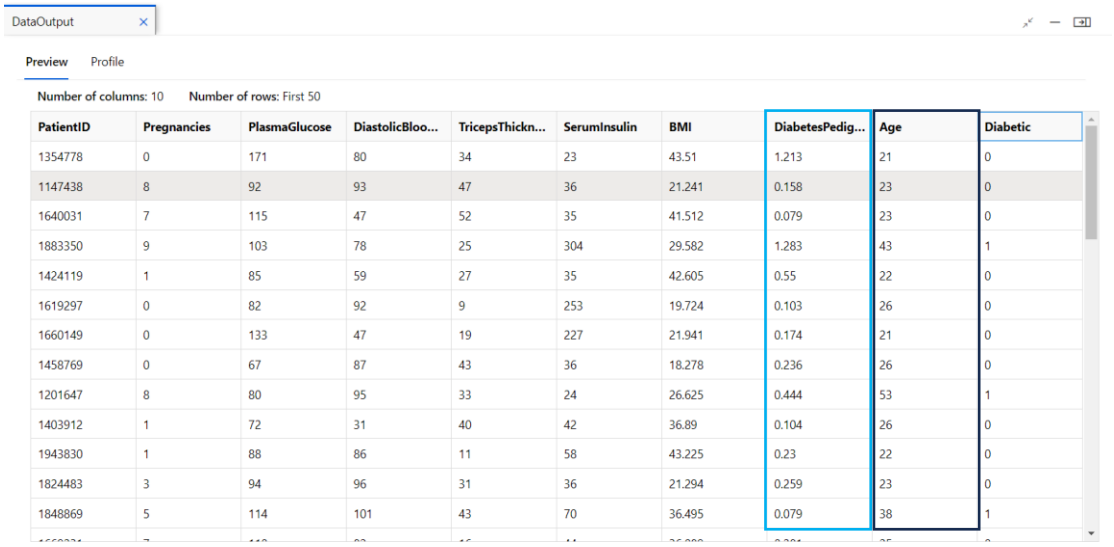


7. Μετακινούμαστε προς τα δεξιά και επιλέγουμε την επικεφαλίδα της στήλης Diabetic και παρατηρούμε ότι περιέχει δύο τιμές 0 και 1. Αυτές οι τιμές αντιπροσωπεύουν τις δύο πιθανές κλάσεις για την ετικέτα που θα προβλέψει το μοντέλο μας, με την τιμή 0 να σημαίνει ότι ο ασθενής δεν έχει διαβήτη και την τιμή 1 να σημαίνει ότι ο ασθενής είναι διαβητικός.



PatientID	Pregnancies	PlasmaGlucose	DiastolicBloo...	TricepsThickn...	SerumInsulin	BMI	DiabetesPedig...	Age	Diabetic
1354778	0	171	80	34	23	43.51	1.213	21	0
1147438	8	92	93	47	36	21.241	0.158	23	0
1640031	7	115	47	52	35	41.512	0.079	23	0
1883350	9	103	78	25	304	29.582	1.283	43	1
1424119	1	85	59	27	35	42.605	0.55	22	0
1619297	0	82	92	9	253	19.724	0.103	26	0
1660149	0	133	47	19	227	21.941	0.174	21	0
1458769	0	67	87	43	36	18.278	0.236	26	0
1201647	8	80	95	33	24	26.625	0.444	53	1
1403912	1	72	31	40	42	36.89	0.104	26	0
1943830	1	88	86	11	58	43.225	0.23	22	0
1824483	3	94	96	31	36	21.294	0.259	23	0
1848869	5	114	101	43	70	36.495	0.079	38	1

8. Κάνουμε κύλιση προς τα αριστερά και επανεξετάζουμε τις άλλες στήλες, οι οποίες αντιπροσωπεύουν τα χαρακτηριστικά που θα χρησιμοποιηθούν για την πρόβλεψη της ετικέτας. Σημειώνουμε ότι οι περισσότερες από αυτές τις στήλες είναι αριθμητικές, αλλά κάθε χαρακτηριστικό έχει τη δική του κλίμακα. Για παράδειγμα, οι τιμές Age κυμαίνονται από 21 έως 77, ενώ οι τιμές DiabetesPedigree κυμαίνονται από 0,078 έως 2,3016. Κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης, είναι μερικές φορές δυνατό οι μεγαλύτερες τιμές να κυριαρχούν στην προκύπτουσα συνάρτηση πρόβλεψης, μειώνοντας την επιρροή των χαρακτηριστικών που βρίσκονται σε μικρότερη κλίμακα. Συνήθως, οι επιστήμονες δεδομένων μετριάζουν αυτή την πιθανή μεροληψία κανονικοποιώντας τις αριθμητικές στήλες ώστε να βρίσκονται σε παρόμοιες κλίμακες.

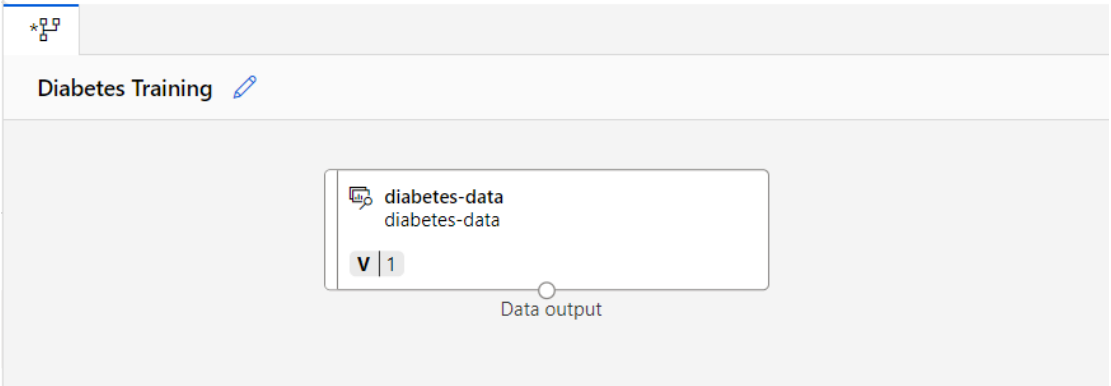

DataOutput

Preview Profile

Number of columns: 10 Number of rows: First 50

PatientID	Pregnancies	PlasmaGlucose	DiastolicBloo...	TricepsThickn...	SerumInsulin	BMI	DiabetesPedig...	Age	Diabetic
1354778	0	171	80	34	23	43.51	1.213	21	0
1147438	8	92	93	47	36	21.241	0.158	23	0
1640031	7	115	47	52	35	41.512	0.079	23	0
1883350	9	103	78	25	304	29.582	1.283	43	1
1424119	1	85	59	27	35	42.605	0.55	22	0
1619297	0	82	92	9	253	19.724	0.103	26	0
1660149	0	133	47	19	227	21.941	0.174	21	0
1458769	0	67	87	43	36	18.278	0.236	26	0
1201647	8	80	95	33	24	26.625	0.444	53	1
1403912	1	72	31	40	42	36.89	0.104	26	0
1943830	1	88	86	11	58	43.225	0.23	22	0
1824483	3	94	96	31	36	21.294	0.259	23	0
1848869	5	114	101	43	70	36.495	0.079	38	1

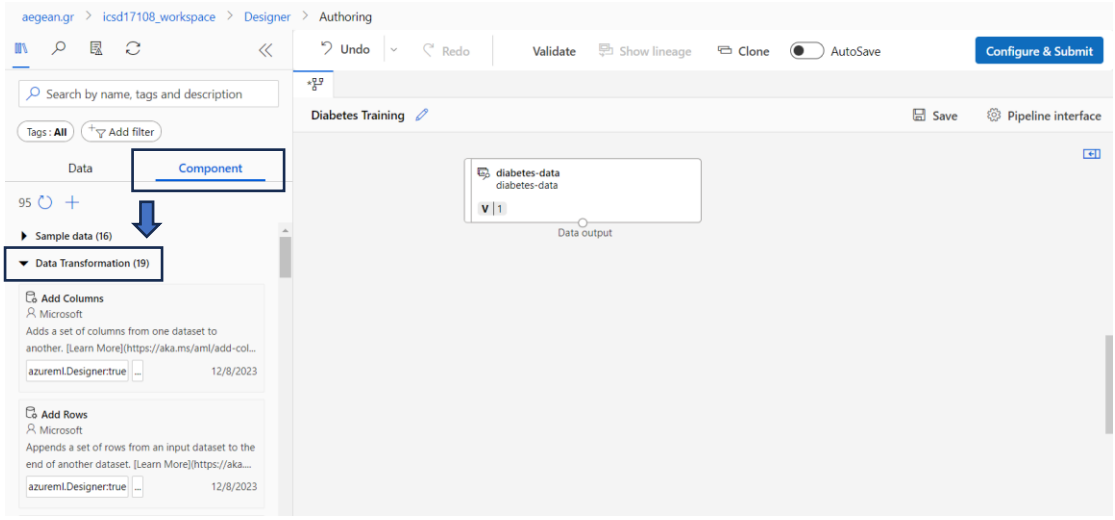
9. Κλείνουμε το παράθυρο απεικόνισης των δεδομένων, ώστε να μπορούμε να δούμε το σύνολο δεδομένων στον καμβά ως εξής:



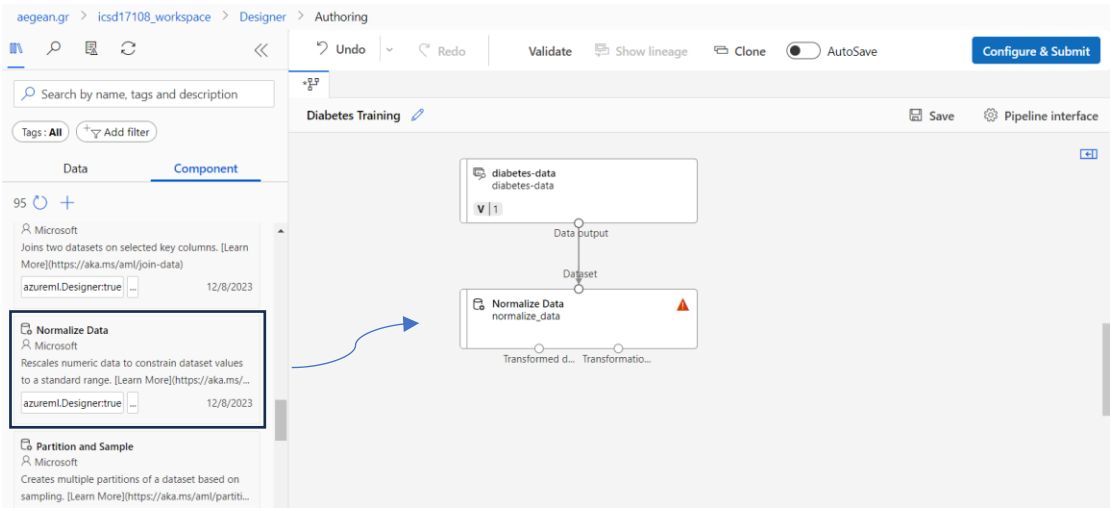
Προσθήκη μετασχηματισμών

Προτού εκπαιδεύσουμε ένα μοντέλο, συνήθως πρέπει να εφαρμόσουμε κάποιους μετασχηματισμούς προ-επεξεργασίας στα δεδομένα.

1. Στο παράθυρο στα αριστερά, αναπτύσσουμε την ενότητα *Data Transformation*, η οποία περιέχει ένα ευρύ φάσμα ενοτήτων που μπορούμε να χρησιμοποιήσουμε για να μετασχηματίσουμε τα δεδομένα πριν από την εκπαίδευση του μοντέλου.

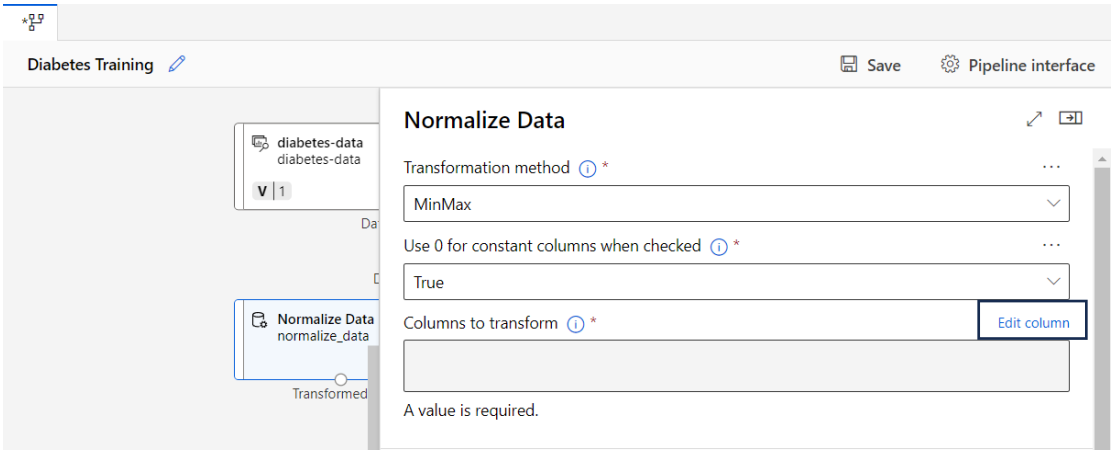


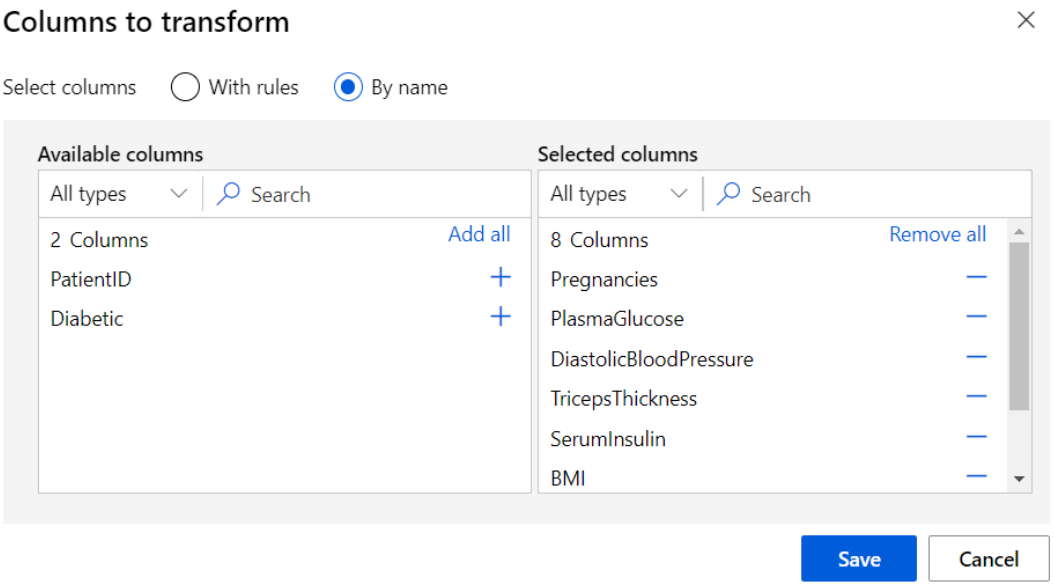
2. Σέρνουμε την ενότητα Normalize Data (Κανονικοποίηση δεδομένων) στον καμβά, κάτω από το σύνολο δεδομένων diabetes-data. Στη συνέχεια, συνδέουμε την έξοδο από το κάτω μέρος του συνόλου δεδομένων diabetes-data με την είσοδο στην κορυφή της ενότητας Normalize Data, όπως παρακάτω:



3. Επιλέγουμε με διπλό κλικ την ενότητα Normalize Data (Κανονικοποίηση δεδομένων) και βλέπουμε τις ρυθμίσεις της, σημειώνοντας ότι απαιτείται να καθορίσουμε τη μέθοδο μετασχηματισμού και τις στήλες που θα μετασχηματιστούν.

4. Ορίζουμε τον μετασχηματισμό σε MinMax και επεξεργαζόμαστε τις στήλες ώστε να συμπεριλάβουμε τις ακόλουθες στήλες ονομαστικά, όπως φαίνεται στην εικόνα:



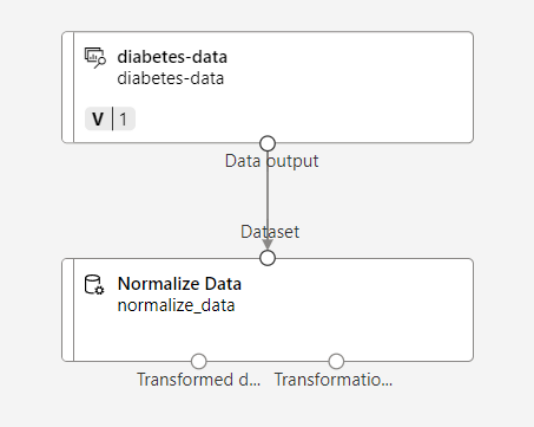


Ο μετασχηματισμός των δεδομένων κανονικοποιεί τις αριθμητικές στήλες για να τις τοποθετήσει στην ίδια κλίμακα, γεγονός που θα βοηθήσει να αποτραπεί η κυριαρχία των στηλών με μεγάλες τιμές στην εκπαίδευση του μοντέλου. Συνήθως θα εφαρμόζαμε ένα σωρό μετασχηματισμούς προ-επεξεργασίας όπως αυτός για να προετοιμάσουμε τα δεδομένα μας για την εκπαίδευση, αλλά θα κρατήσουμε τα πράγματα απλά σε αυτή το σενάριο.

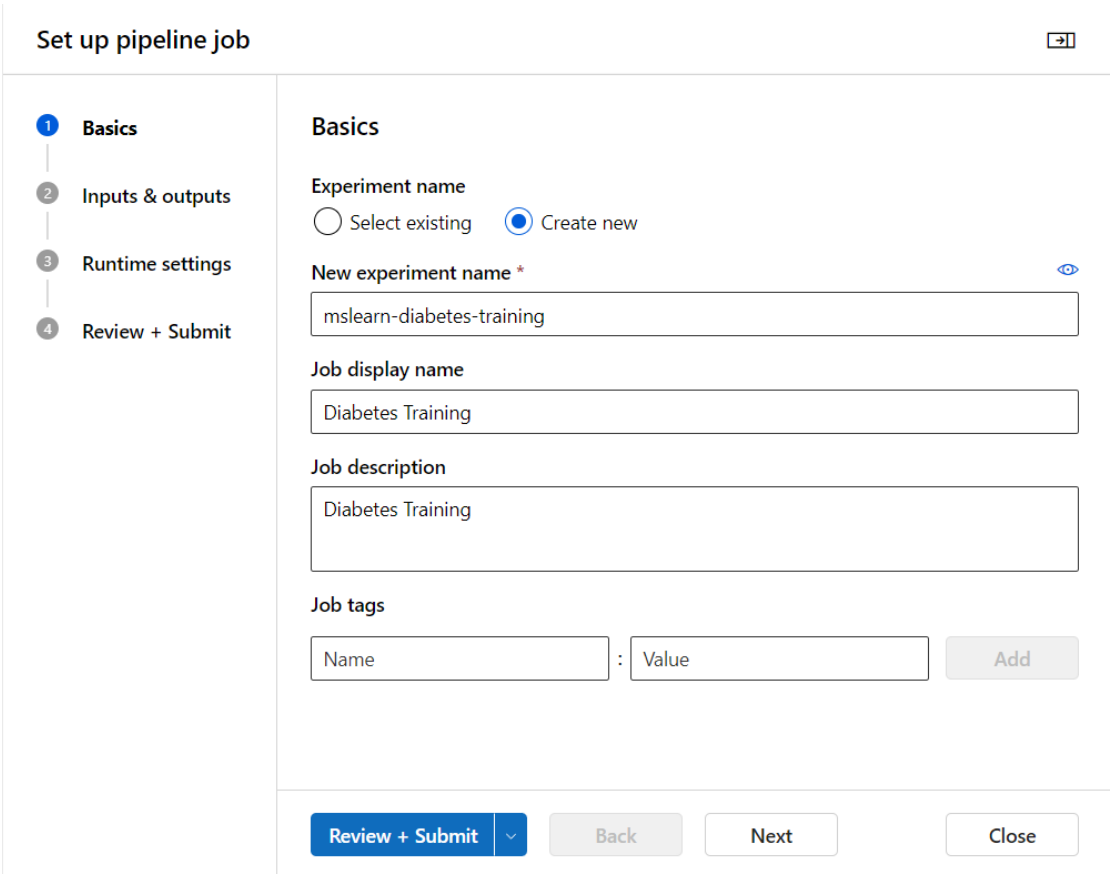
Εκτέλεση του αγωγού

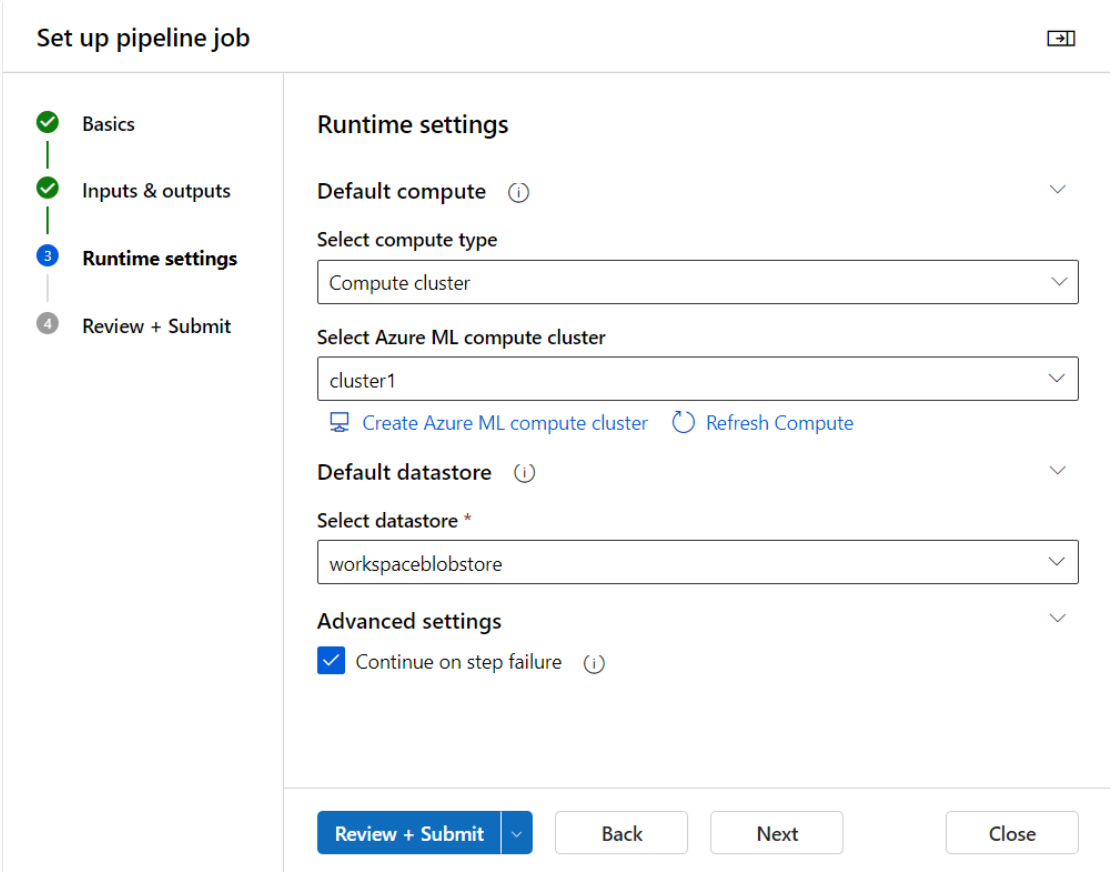
Για να εφαρμόσουμε τους μετασχηματισμούς των δεδομένων μας, πρέπει να εκτελέσουμε τον αγωγό ως πείραμα.

1. Βεβαιωνόμαστε ότι ο αγωγός μας μοιάζει με αυτό:



2. Επιλέγουμε Configure & Submit και εκτελούμε τον αγωγό ως νέο πείραμα με όνομα mslearn-diabetes-training στη συστάδα υπολογιστών μας, πατώντας Submit.



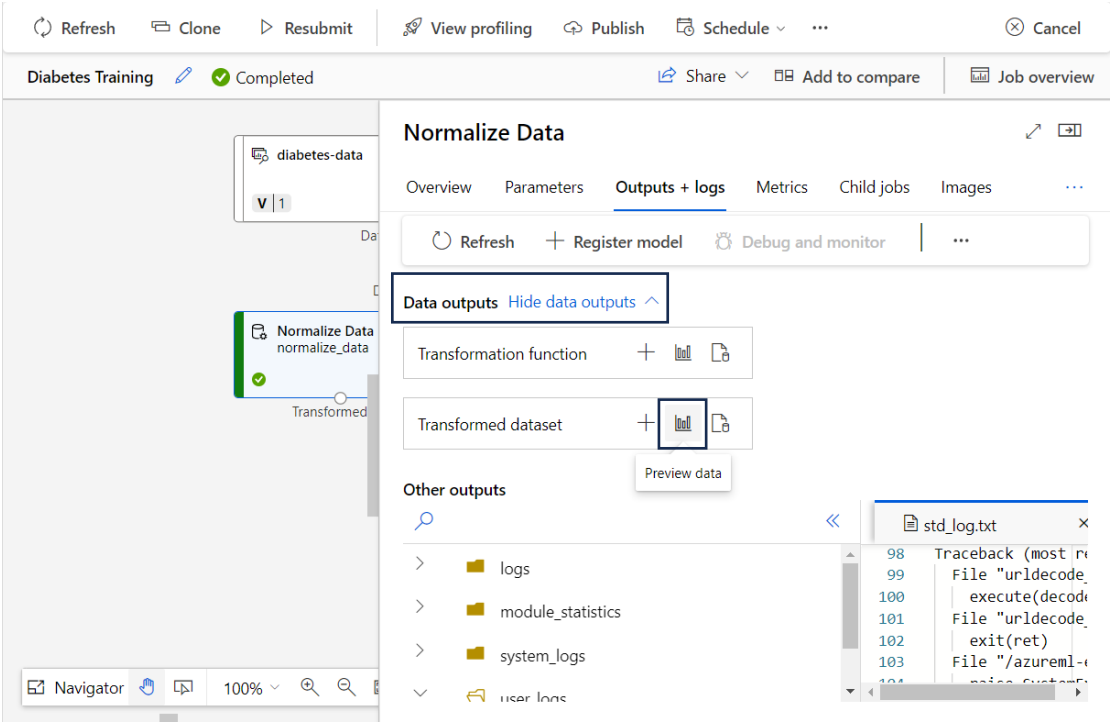


3. Περιμένουμε να ολοκληρωθεί η εκτέλεση, η οποία μπορεί να διαρκέσει μερικά λεπτά.

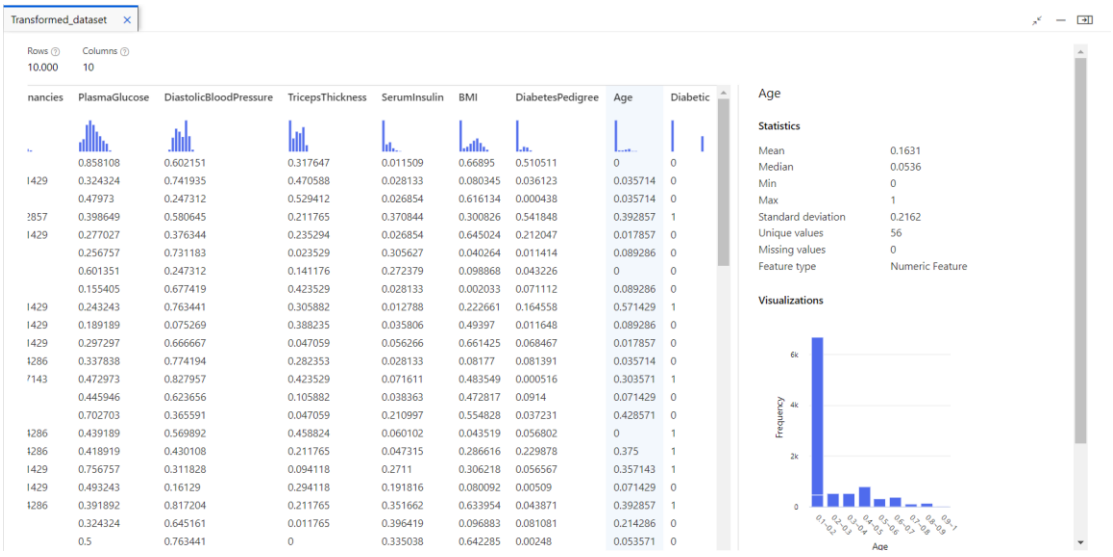
Προβολή των μετασχηματισμένων δεδομένων

Το σύνολο δεδομένων είναι τώρα έτοιμο για την εκπαίδευση του μοντέλου.

1. Επιλέγουμε με διπλό κλικ την ολοκληρωμένη ενότητα Normalize Data (Κανονικοποίηση δεδομένων) και στο παράθυρο ρυθμίσεων στα δεξιά της, στην καρτέλα Outputs + logs (Έξοδοι + αρχεία καταγραφής), επιλέγουμε το εικονίδιο Preview Data για το Transformed dataset (Μετασχηματισμένο σύνολο δεδομένων).



2. Προβάλουμε τα δεδομένα, σημειώνοντας ότι οι αριθμητικές στήλες που επιλέξαμε έχουν κανονικοποιηθεί σε κοινή κλίμακα.



3. Κλείνουμε την οπτικοποίηση των κανονικοποιημένων δεδομένων.

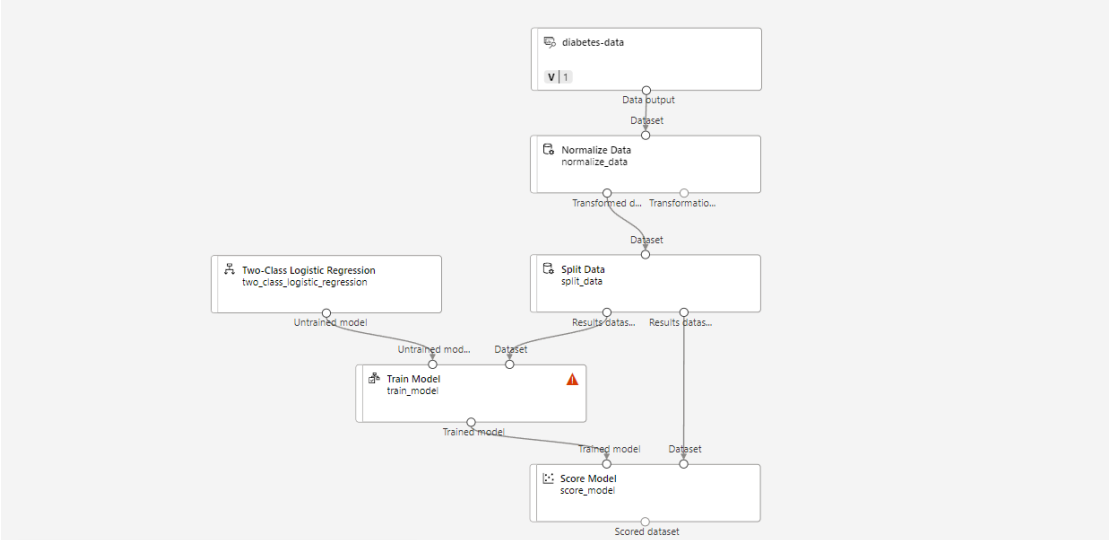
4.3.4 Δημιουργία και εκτέλεση ενός αγωγού κατάρτισης

Αφού χρησιμοποιήσαμε μετασχηματισμούς δεδομένων για να προετοιμάσουμε τα δεδομένα, μπορούμε να τα χρησιμοποιήσουμε για να εκπαιδεύσουμε ένα μοντέλο μηχανικής μάθησης.

Προσθήκη ενοτήτων εκπαίδευσης

Είναι κοινή πρακτική να εκπαιδεύεται το μοντέλο χρησιμοποιώντας ένα υποσύνολο των δεδομένων, κρατώντας παράλληλα κάποια δεδομένα με τα οποία θα δοκιμαστεί το εκπαιδευμένο μοντέλο. Αυτό μας επιτρέπει να συγκρίνουμε τις ετικέτες που προβλέπει το μοντέλο με τις πραγματικές γνωστές ετικέτες στο αρχικό σύνολο δεδομένων.

Σε αυτό το σενάριο, θα επεκτείνουμε την αγωγό εκπαίδευσης για τον διαβήτη, όπως φαίνεται παρακάτω:



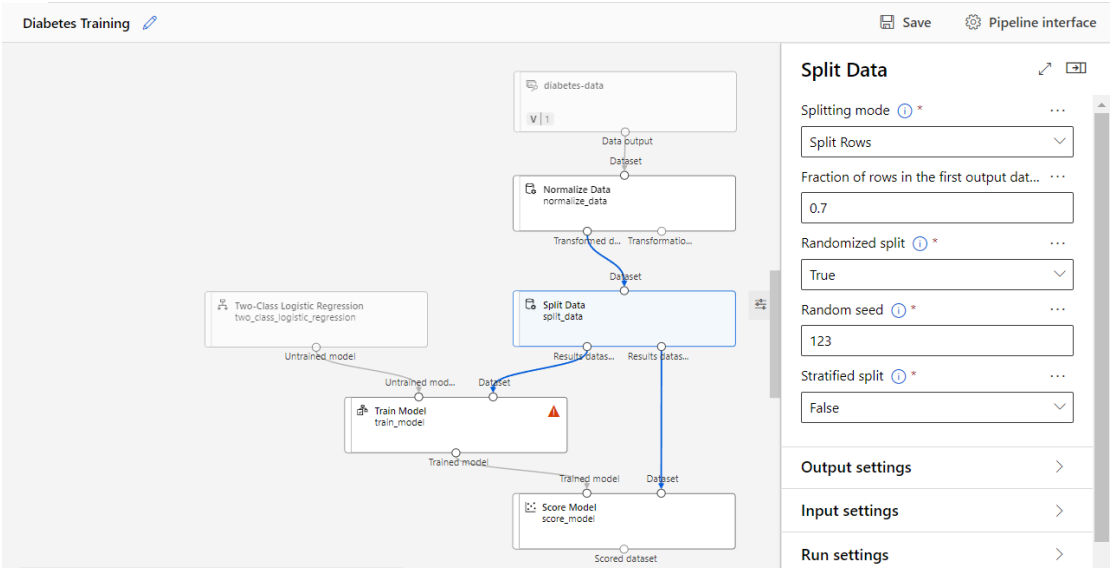
Ακολουθούμε τα παρακάτω βήματα, χρησιμοποιώντας την παραπάνω εικόνα ως αναφορά καθώς προσθέτουμε και ρυθμίζουμε τις απαιτούμενες ενότητες.

1. Ανοίγουμε τον αγωγό εκπαίδευσης για τον διαβήτη που δημιουργήσαμε στην προηγούμενη ενότητα, αν δεν είναι ήδη ανοιχτός.

2. Στο παράθυρο στα αριστερά, στην ενότητα Data Transformations (Μετασχηματισμοί δεδομένων), σέρνουμε μια ενότητα Split Data (Διαχωρισμός δεδομένων) στον καμβά κάτω από την ενότητα Normalize Data (Κανονικοποίηση δεδομένων). Στη συνέχεια, συνδέουμε την έξοδο Transformed Dataset (αριστερά) της μονάδας Normalize Data (Κανονικοποίηση δεδομένων) με την είσοδο της μονάδας Split Data (Διαχωρισμός δεδομένων).

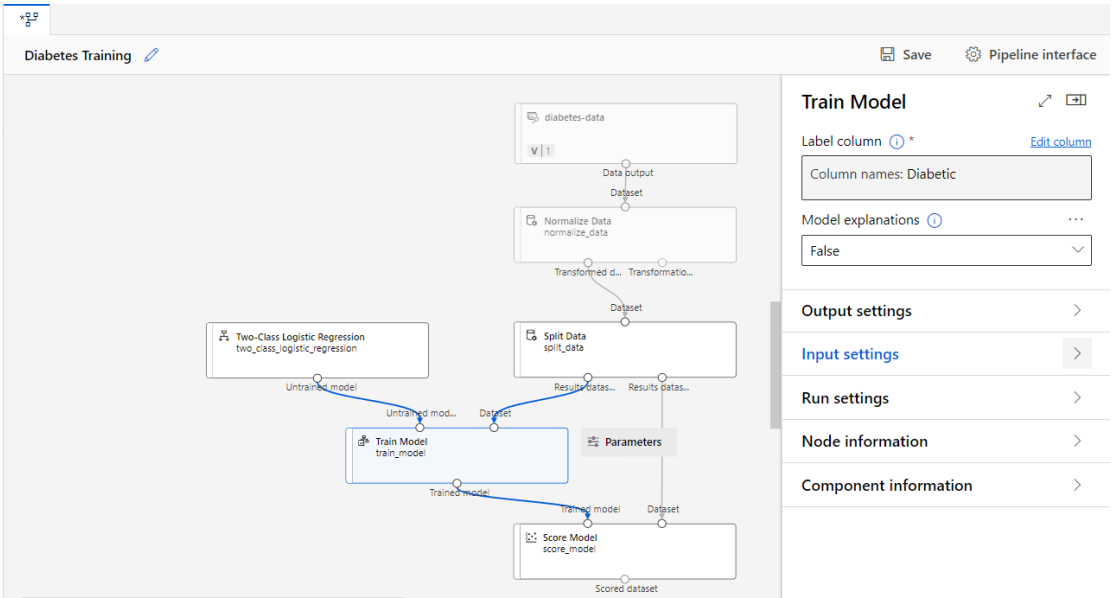
3. Επιλέγουμε την ενότητα Split Data και διαμορφώνουμε τις ρυθμίσεις της ως εξής:

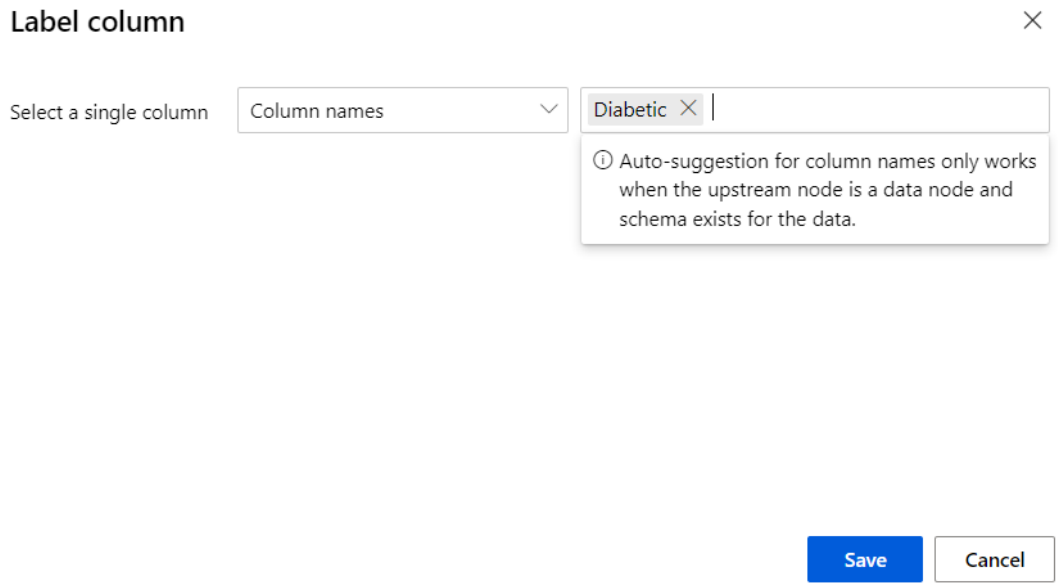
- **Splitting mode**: Split Rows
- **Fraction of rows in the first output dataset**: 0.7
- **Random seed**: 123
- **Stratified split**: False



4. Αναπτύσσουμε την ενότητα Model Training (Εκπαίδευση μοντέλου) στο παράθυρο στα αριστερά και σέρνουμε μια ενότητα Εκπαίδευση μοντέλου στον καμβά, κάτω από την ενότητα Διαχωρισμός δεδομένων. Στη συνέχεια, συνδέουμε την έξοδο Result dataset1 (αριστερά) της μονάδας Split Data (Διαχωρισμένα δεδομένα) με την είσοδο Dataset (δεξιά) της μονάδας Train Model (Εκπαίδευση μοντέλου).

5. Το μοντέλο που εκπαιδεύουμε θα προβλέψει την τιμή Diabetic, οπότε επιλέγουμε το Train Model module και τροποποιούμε τις ρυθμίσεις του ώστε να ορίσουμε τη στήλη Label σε Diabetic (ταιριάζοντας ακριβώς με την περίπτωση και την ορθογραφία!)

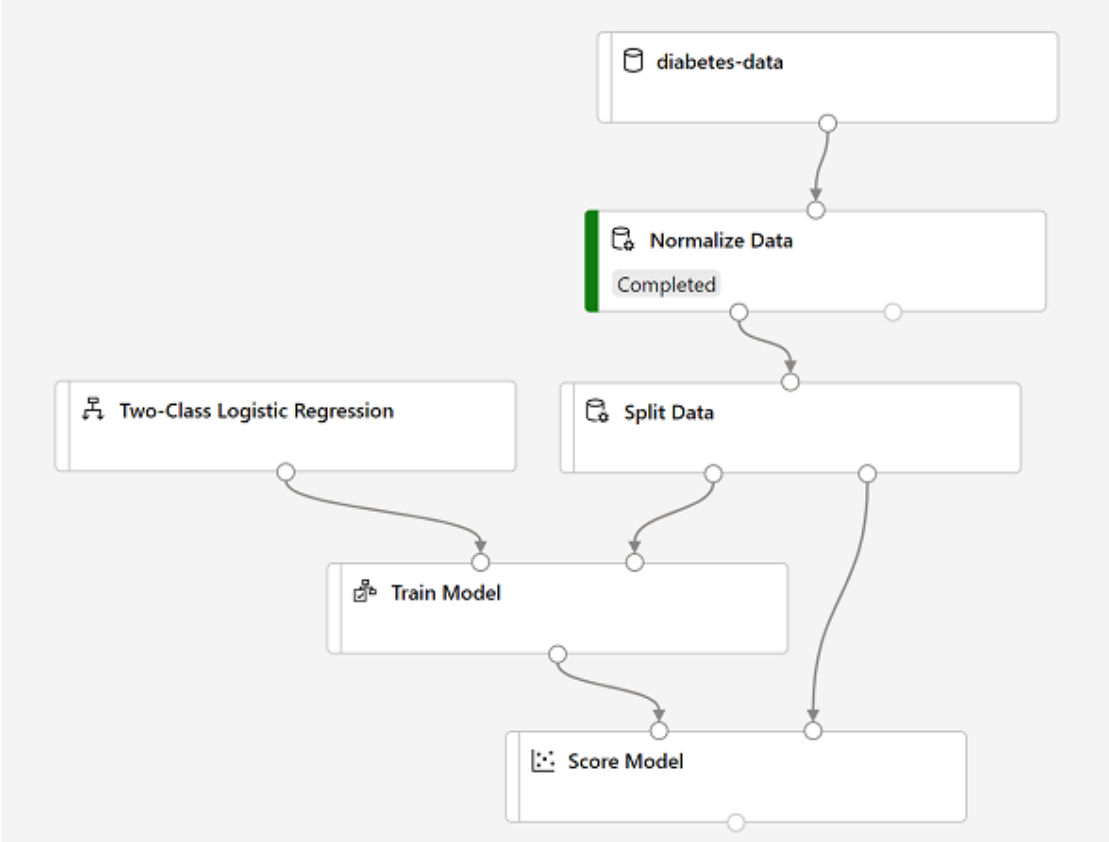




6. Η ετικέτα Diabetic που θα προβλέψει το μοντέλο είναι μια κλάση (0 ή 1), οπότε πρέπει να εκπαιδεύσουμε το μοντέλο χρησιμοποιώντας έναν αλγόριθμο ταξινόμησης. Συγκεκριμένα, υπάρχουν δύο πιθανές κλάσεις, οπότε χρειαζόμαστε έναν δυαδικό αλγόριθμο ταξινόμησης. Αναπτύσσουμε την ενότητα Αλγόριθμοι μηχανικής μάθησης και στην ενότητα Ταξινόμηση, σέρνουμε μια ενότητα λογιστικής παλινδρόμησης δύο κλάσεων στον καμβά, στα αριστερά της ενότητας Διαχωρισμός δεδομένων και πάνω από την ενότητα Εκπαίδευση μοντέλου. Στη συνέχεια, συνδέουμε την έξοδό της στην είσοδο Untrained model (αριστερά) της ενότητας Train Model.

7. Για να δοκιμάσουμε το εκπαιδευμένο μοντέλο, πρέπει να το χρησιμοποιήσουμε για να βαθμολογήσουμε το σύνολο δεδομένων επικύρωσης που κρατήσαμε πίσω όταν χωρίσαμε τα αρχικά δεδομένα. Με άλλα λόγια, να προβλέψουμε ετικέτες για τα χαρακτηριστικά στο σύνολο δεδομένων επικύρωσης. Αναπτύσσουμε την ενότητα Model Scoring & Evaluation (Βαθμολόγηση και αξιολόγηση μοντέλου) και σέρνουμε μια ενότητα Score Model (Βαθμολόγηση μοντέλου) στον καμβά, κάτω από την ενότητα Train Model (Εκπαίδευση μοντέλου). Στη συνέχεια, συνδέουμε την έξοδο της ενότητας Train Model με την είσοδο Trained model (αριστερά) της ενότητας Score Model και συνδέουμε την έξοδο Results dataset2 (δεξιά) της ενότητας Split Data με την είσοδο Dataset (δεξιά) της ενότητας Score Model.

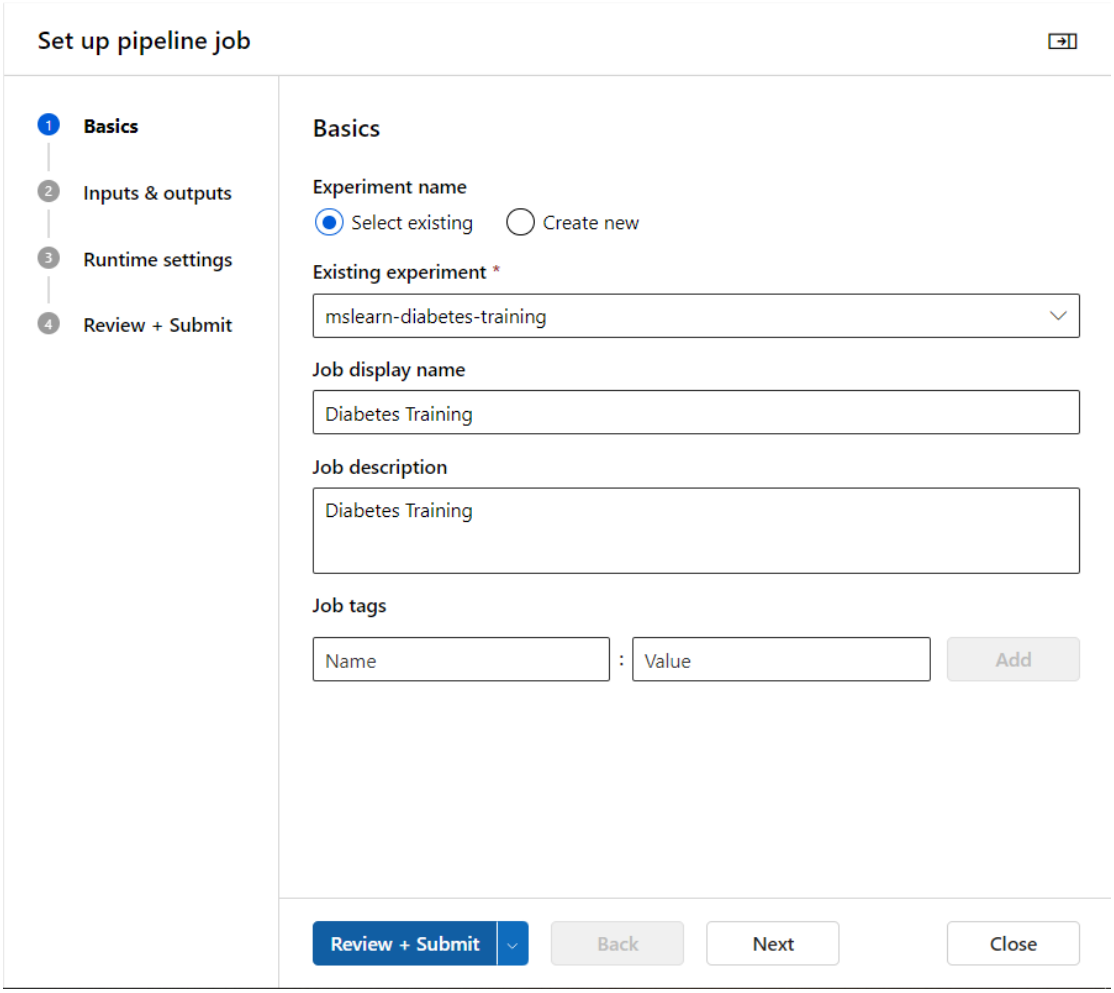
8. Βεβαιωνόμαστε πως ο αγωγός μας έχει την εξής μορφή:



Εκτέλεση εκπαιδευτικού αγωγού

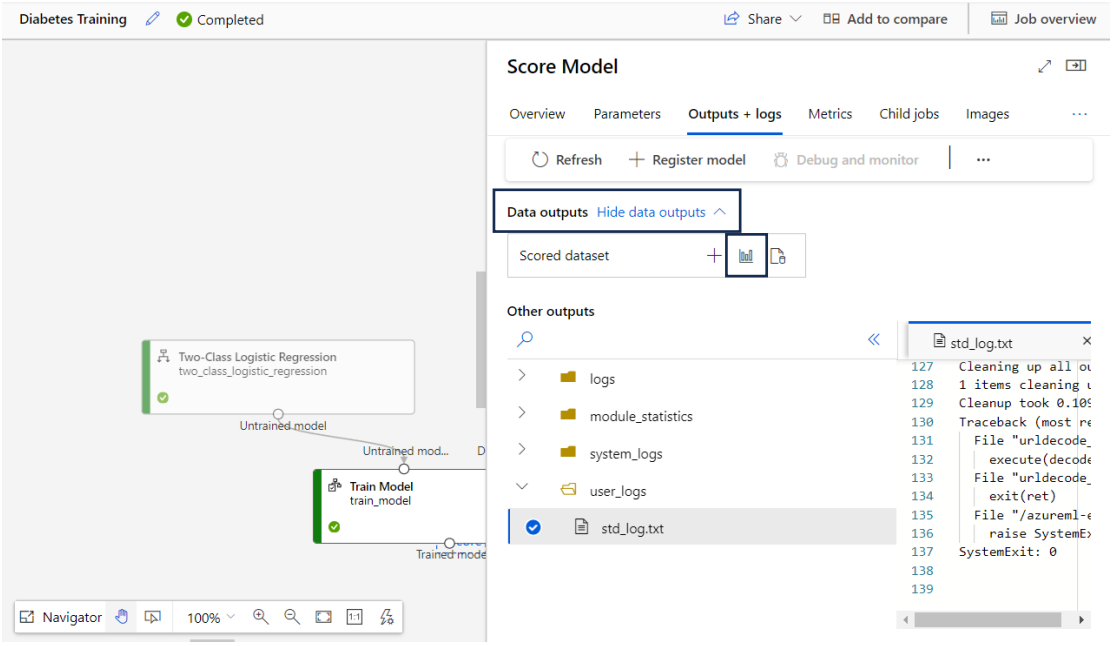
Τώρα είμαστε έτοιμοι να εκτελέσουμε τον αγωγό εκπαίδευσης και να εκπαιδεύσουμε το μοντέλο.

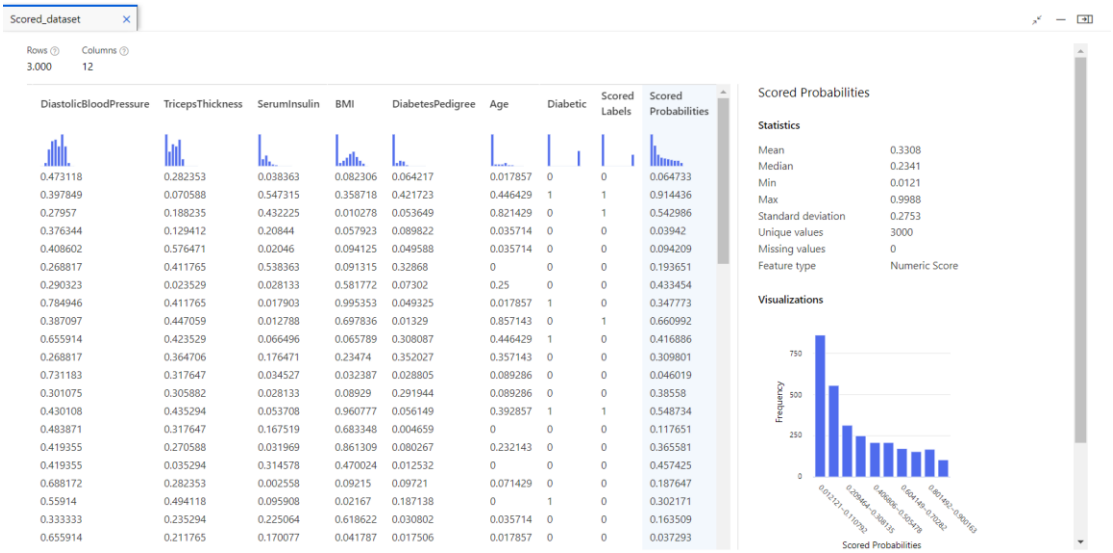
1. Επιλέγουμε Submit (Υποβολή) και εκτελούμε τον αγωγό εκπαίδευσης χρησιμοποιώντας το υπάρχον πείραμα με όνομα mslearn-diabetes-training.



2. Περιμένουμε να ολοκληρωθεί η εκτέλεση του πειράματος, η οποία μπορεί να διαρκέσει 5 λεπτά ή και περισσότερο.

3. Όταν ολοκληρωθεί η εκτέλεση του πειράματος, επιλέγουμε την ενότητα Score Model και στο παράθυρο ρυθμίσεων, στην καρτέλα Outputs + Logs (Έξοδοι + αρχεία καταγραφής), στην ενότητα Data outputs (Έξοδοι δεδομένων), στην ενότητα Scored dataset (Σύνολο δεδομένων με βαθμολογία), χρησιμοποιούμε το εικονίδιο Visualize (Οπτικοποίηση) για να προβάλουμε τα αποτελέσματα.





4. Μετακινούμαστε προς τα δεξιά και παρατηρούμε ότι δίπλα στη στήλη Diabetic (η οποία περιέχει τις γνωστές πραγματικές τιμές της ετικέτας) υπάρχει μια νέα στήλη με το όνομα Scored Labels, η οποία περιέχει τις προβλεπόμενες τιμές της ετικέτας, και μια στήλη Scored Probabilities που περιέχει μια τιμή πιθανότητας μεταξύ 0 και 1. Αυτή υποδηλώνει την πιθανότητα μιας θετικής πρόβλεψης, έτσι ώστε πιθανότητες μεγαλύτερες από 0,5 οδηγούν σε μια προβλεπόμενη ετικέτα 1 (διαβητικός), ενώ πιθανότητες μεταξύ 0 και 0,5 οδηγούν σε μια προβλεπόμενη ετικέτα 0 (όχι διαβητικός).

Κλείνουμε το παράθυρο απεικόνισης αποτελεσμάτων του μοντέλου βαθμολογίας. Το μοντέλο προβλέπει τιμές για την ετικέτα Diabetic, αλλά πόσο αξιόπιστες είναι οι προβλέψεις του; Για να το εκτιμήσουμε αυτό, πρέπει να αξιολογήσουμε το μοντέλο.

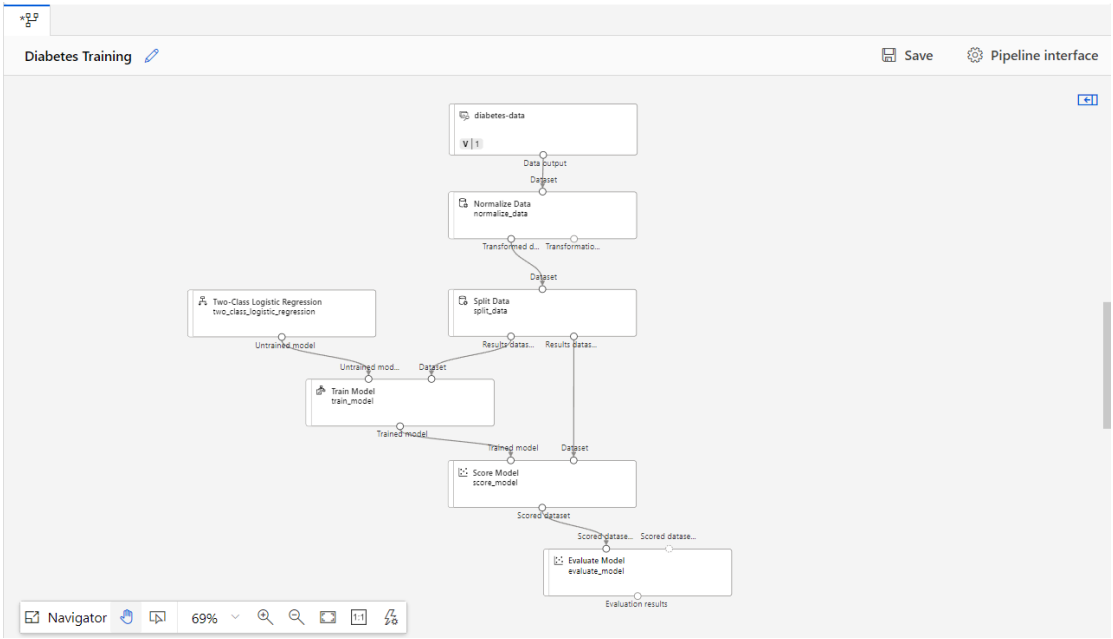
4.3.5 Αξιολόγηση ενός μοντέλου ταξινόμησης

Τα δεδομένα επικύρωσης που συγκρατήσαμε και χρησιμοποιήσαμε για τη βαθμολόγηση του μοντέλου περιλαμβάνουν τις γνωστές τιμές για την ετικέτα. Έτσι, για να επικυρώσουμε το μοντέλο, μπορούμε να συγκρίνουμε τις πραγματικές τιμές για την ετικέτα με τις τιμές της ετικέτας που προβλέφθηκαν όταν βαθμολογήσαμε το σύνολο δεδομένων επικύρωσης. Με βάση αυτή τη σύγκριση, μπορούμε να υπολογίσουμε διάφορες τιμές που περιγράφουν πόσο καλά αποδίδει το μοντέλο.

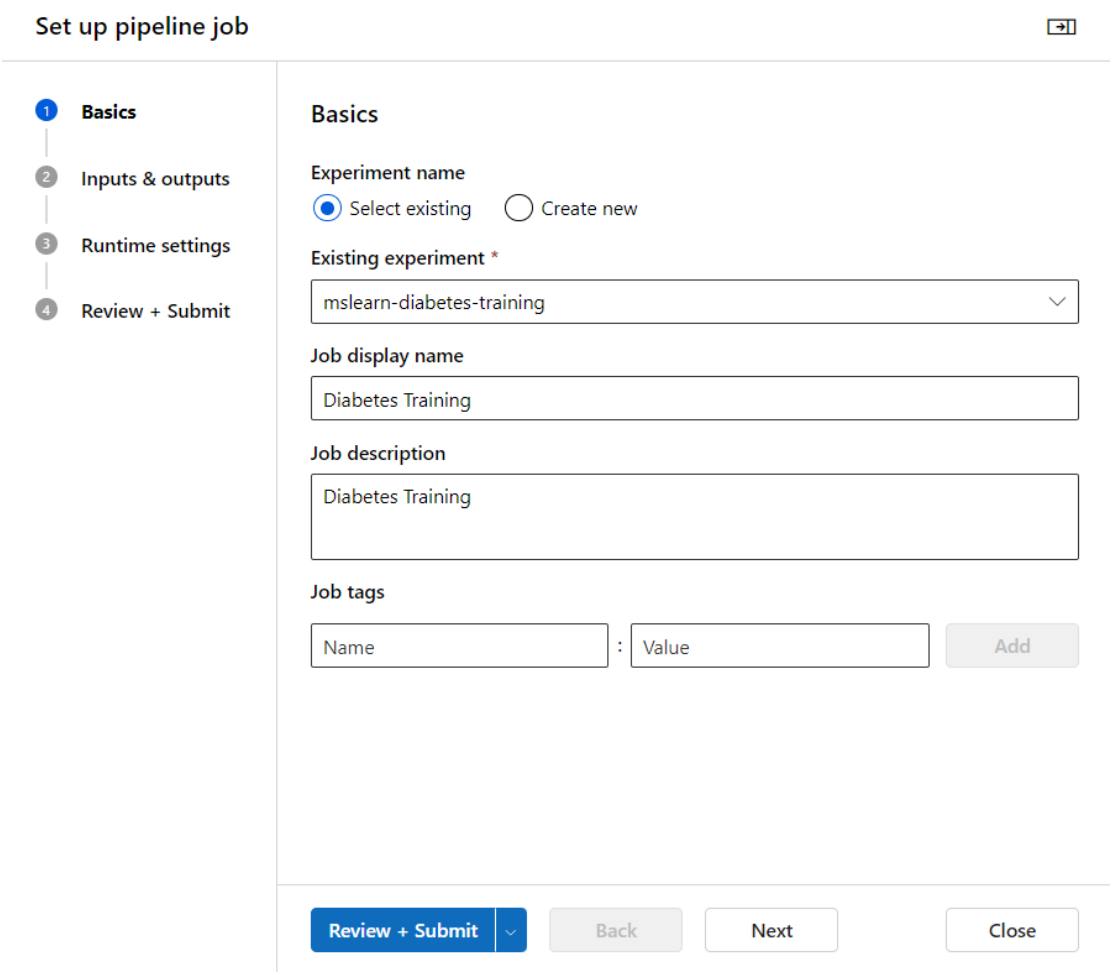
Προσθήκη ενότητας Evaluate Model (Αξιολόγηση μοντέλου)

1. Ανοίγουμε τον αγωγό Diabetes Training που δημιουργήσαμε στην προηγούμενη ενότητα, αν δεν είναι ήδη ανοιχτός.

2. Στο παράθυρο στα αριστερά, στην ενότητα Model Scoring & Evaluation, σέρνουμε μια ενότητα Evaluate Model στον καμβά, κάτω από την ενότητα Score Model, και συνδέουμε την έξοδο της ενότητας Score Model με την είσοδο Scored dataset (αριστερά) της ενότητας Evaluate Model.

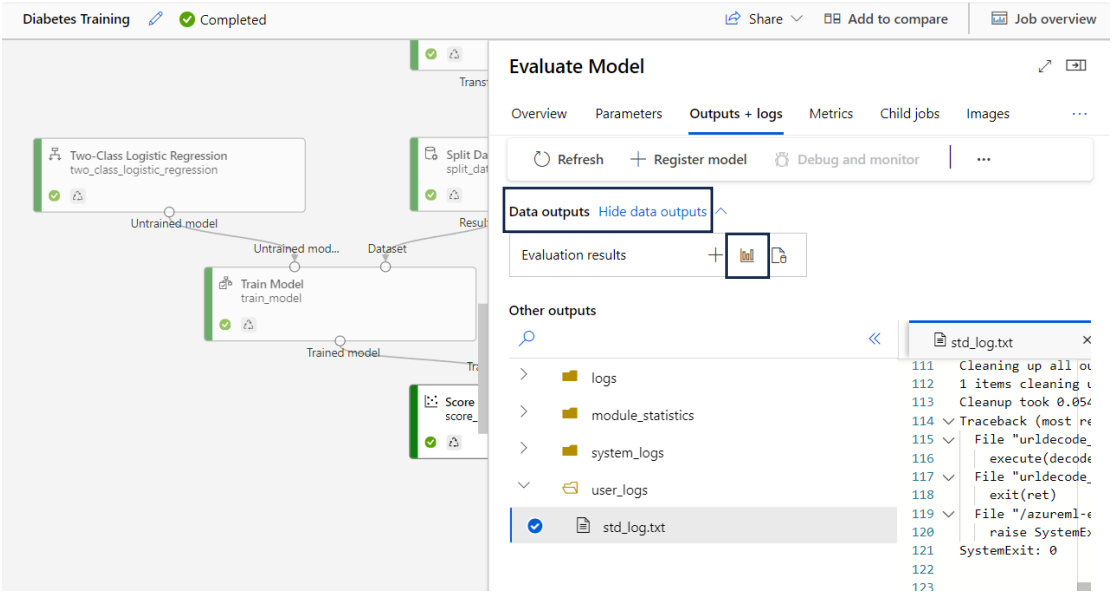


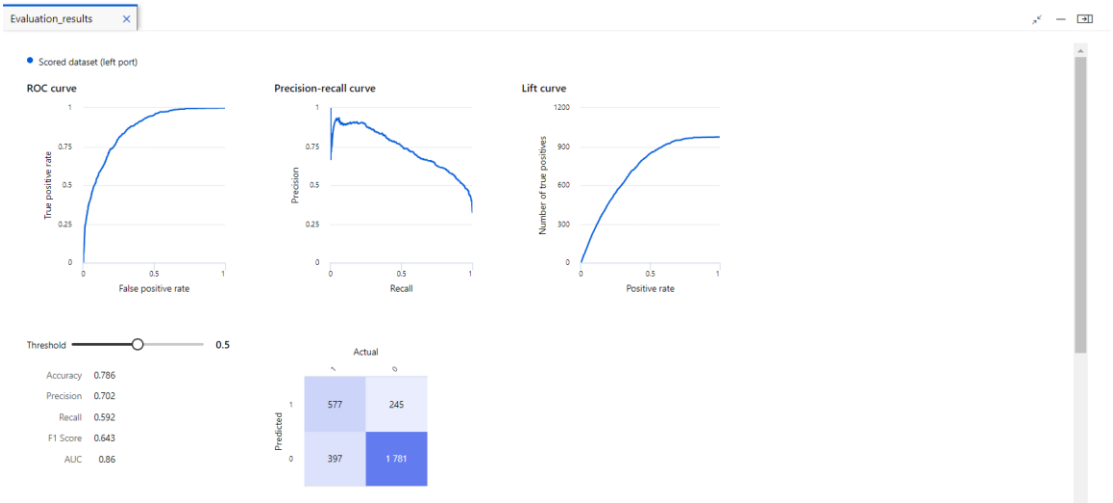
4. Επιλέγουμε Submit και εκτελούμε τον αγωγό χρησιμοποιώντας το υπάρχον πείραμα με όνομα mslearn-diabetes-training.



5. Περιμένουμε να ολοκληρωθεί η εκτέλεση του πειράματος.

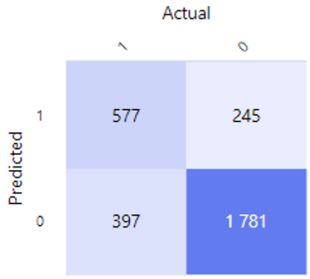
6. Όταν ολοκληρωθεί η εκτέλεση του πειράματος, επιλέγουμε την ενότητα Evaluate Model (Αξιολόγηση μοντέλου) και στο παράθυρο ρυθμίσεων, στην καρτέλα Outputs + Logs (Έξοδοι + αρχεία καταγραφής), στην ενότητα Data outputs (Έξοδοι δεδομένων) στην ενότητα Evaluation results (Αποτελέσματα αξιολόγησης), χρησιμοποιούμε το εικονίδιο Visualize (Οπτικοποίηση) για να προβάλουμε τις μετρήσεις απόδοσης. Αυτές οι μετρήσεις μπορούν να βοηθήσουν τους επιστήμονες δεδομένων να αξιολογήσουν πόσο καλά προβλέπει το μοντέλο με βάση τα δεδομένα επικύρωσης.



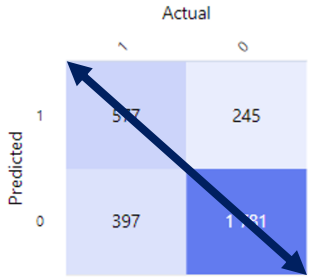


Score bin	Positive exam...	Negative exam...	Fraction above thres...	Accura...	F1 Score	Precisi...	Recall	Negative precis...	Negative recall	Cumulative AUC
(0.900,1.000]	95	11	0.035	0.703	0.176	0.896	0.098	0.696	0.995	0.000
(0.800,0.900]	149	21	0.092	0.746	0.390	0.884	0.251	0.732	0.984	0.002
(0.700,0.800]	116	45	0.146	0.770	0.510	0.824	0.370	0.760	0.962	0.009
(0.600,0.700]	108	65	0.203	0.784	0.591	0.767	0.480	0.788	0.930	0.023
(0.500,0.600]	109	103	0.274	0.786	0.643	0.702	0.592	0.818	0.879	0.050
(0.400,0.500]	110	111	0.348	0.786	0.681	0.659	0.705	0.853	0.824	0.086
(0.300,0.400]	99	153	0.432	0.768	0.693	0.607	0.807	0.890	0.749	0.142
(0.200,0.300]	88	242	0.542	0.716	0.673	0.538	0.897	0.927	0.629	0.245
(0.100,0.200]	85	517	0.742	0.572	0.599	0.431	0.985	0.981	0.374	0.487
(0.000,0.100]	15	758	1.000	0.325	0.490	0.325	1.000	1.000	0.000	0.860

Ο πίνακας σύγχυσης δείχνει τις περιπτώσεις όπου τόσο η προβλεπόμενη όσο και η πραγματική τιμή ήταν 1 (γνωστές ως αληθώς θετικές) πάνω αριστερά και τις περιπτώσεις όπου τόσο η προβλεπόμενη όσο και η πραγματική τιμή ήταν 0 (αληθώς αρνητικές) κάτω δεξιά. Τα άλλα κελιά δείχνουν περιπτώσεις όπου οι προβλεπόμενες και οι πραγματικές τιμές διαφέρουν (ψευδώς θετικά και ψευδώς αρνητικά).



Τα κελιά του πίνακα είναι χρωματισμένα έτσι ώστε όσο περισσότερες περιπτώσεις αντιπροσωπεύονται στο κελί, τόσο πιο έντονο είναι το χρώμα, με αποτέλεσμα να μπορούμε να εντοπίσουμε ένα μοντέλο που προβλέπει με ακρίβεια για όλες τις κλάσεις αναζητώντας μια διαγώνια γραμμή από έντονα χρωματισμένα κελιά από πάνω αριστερά προς κάτω δεξιά. Με άλλα λόγια, τα κελιά όπου οι προβλεπόμενες τιμές ταιριάζουν με τις πραγματικές τιμές.



Για ένα μοντέλο ταξινόμησης πολλαπλών κλάσεων, όπου υπάρχουν περισσότερες από δύο πιθανές κλάσεις, η ίδια προσέγγιση χρησιμοποιείται για την καταγραφή σε πίνακες κάθε πιθανού συνδυασμού των πραγματικών και προβλεπόμενων τιμών. Έτσι ένα μοντέλο με τρεις πιθανές κλάσεις θα έχει ως αποτέλεσμα έναν πίνακα 3x3 με μια διαγώνια γραμμή κελιών όπου οι προβλεπόμενες και οι πραγματικές ετικέτες ταιριάζουν.

8. Επανεξετάζουμε τις μετρήσεις στα αριστερά του πίνακα σύγχυσης, οι οποίες περιλαμβάνουν:

- **Accuracy (Ακρίβεια)**: Ο λόγος των σωστών προβλέψεων (αληθώς θετικές + αληθώς αρνητικές) προς τον συνολικό αριθμό των προβλέψεων. Με άλλα λόγια, το ποσοστό των προβλέψεων για τον διαβήτη, σύμφωνα με τον πίνακα, υπολογίζεται ως εξής:

$$\frac{577 + 1781}{577 + 245 + 397 + 1781} = \frac{2358}{3000} = 0.786$$

- **Precision (Ακρίβεια)**: Το κλάσμα των θετικών περιπτώσεων που αναγνωρίστηκαν σωστά, δηλαδή ο αριθμός των αληθώς θετικών (1,1) δια του αριθμού των αληθώς θετικών (1,1) συν των ψευδώς θετικών (1,0). Με άλλα λόγια, από όλους τους ασθενείς που το μοντέλο προέβλεψε ότι έχουν διαβήτη, το ποσοστό του πόσοι είναι πραγματικά διαβητικοί, υπολογίζεται ως εξής:

$$\frac{577}{577 + 245} = \frac{577}{822} = 0.702$$

- **Recall (Ανάκληση)**: Το κλάσμα των περιπτώσεων που ταξινομούνται ως θετικές και είναι πράγματι θετικές, δηλαδή ο αριθμός των αληθώς θετικών (1,1) διαιρείται με τον αριθμό των αληθώς θετικών (1,1) συν τα ψευδώς θετικά (1,0). Με άλλα λόγια, από όλους τους ασθενείς που έχουν πράγματι διαβήτη, πόσους αναγνώρισε το μοντέλο;

$$\frac{577}{577 + 397} = \frac{577}{974} = 0.592$$

- **F1 Score (Βαθμολογία F1)**: Μια συνολική μέτρηση που ουσιαστικά συνδυάζει την ακρίβεια και την ανάκληση.
- **AUC (Περιοχή κάτω από την καμπύλη ROC)**: Η AUC μετρά ολόκληρη τη δισδιάστατη περιοχή κάτω από την καμπύλη ROC, σαν τον ολοκληρωτικό λογισμό, από το (0,0) έως το (1,1).

Από αυτές τις μετρήσεις, η ακρίβεια (accuracy) είναι η πιο διαισθητική. Ωστόσο, θα πρέπει να είμαστε προσεκτικοί όσον αφορά τη χρήση της απλής ακρίβειας ως μέτρο του πόσο καλά λειτουργεί ένα μοντέλο. Ας υποθέσουμε ότι μόνο το 3% του πληθυσμού είναι διαβητικοί. Θα μπορούσαμε να δημιουργήσουμε ένα μοντέλο που προβλέπει πάντα το 0 και θα ήταν 97% ακριβές, απλώς δεν θα ήταν πολύ χρήσιμο! Για το λόγο αυτό, οι περισσότεροι επιστήμονες δεδομένων χρησιμοποιούν άλλες μετρικές όπως η ακρίβεια και η ανάκληση για να αξιολογήσουν την απόδοση του μοντέλου ταξινόμησης.

9. Πάνω από τη λίστα των μετρικών, παρατηρούμε ότι υπάρχει ένα ρυθμιστικό κατώφλι (Threshold). Έχουμε πάντα υπόψιν ότι αυτό που προβλέπει ένα μοντέλο ταξινόμησης είναι η πιθανότητα για κάθε πιθανή κλάση.

Στην περίπτωση αυτού του δυαδικού μοντέλου ταξινόμησης, η προβλεπόμενη πιθανότητα για μια θετική πρόβλεψη (δηλαδή διαβήτη) είναι μια τιμή μεταξύ 0 και 1. Εξ ορισμού, μια προβλεπόμενη πιθανότητα για διαβήτη πάνω από 0,5 οδηγεί σε μια πρόβλεψη κλάσης 1, ενώ μια πρόβλεψη κάτω από αυτό το όριο σημαίνει ότι υπάρχει μεγαλύτερη πιθανότητα ο ασθενής να μην έχει διαβήτη (οι πιθανότητες για όλες τις κλάσεις αθροίζονται στο 1), οπότε η προβλεπόμενη κλάση θα είναι 0.

Σε περίπτωση που μετακινήσουμε το ρυθμιστικό του κατωφλίου τότε παρατηρούμε την επίδραση στον πίνακα σύγχυσης. Αν το μετακινήσουμε τελείως προς τα αριστερά (0), η μετρική ανάκληση γίνεται 1, και αν το μετακινήσουμε τελείως προς τα δεξιά (1), η μετρική ανάκληση γίνεται 0.

Η καμπύλη ROC (receiver operating characteristic curve) είναι ένα γραφικό διάγραμμα που απεικονίζει την απόδοση ενός μοντέλου δυαδικής ταξινόμησης, που μπορεί να χρησιμοποιηθεί και για ταξινόμηση πολλαπλών κλάσεων, σε διαφορετικές τιμές κατωφλίου (Threshold). Πρακτικά, η καμπύλη ROC είναι η γραφική παράσταση του πραγματικού θετικού ποσοστού (TPR - True positive rate) έναντι του ψευδώς θετικού ποσοστού (FPR - False positive rate) σε κάθε ρύθμιση κατωφλίου), η οποία μετρά τον αριθμό των αρνητικών περιπτώσεων που αναγνωρίστηκαν εσφαλμένα ως θετικές σε σύγκριση με τον αριθμό των πραγματικών αρνητικών περιπτώσεων.

10. Η απεικόνιση αυτών των μετρικών μεταξύ τους για κάθε πιθανή τιμή κατωφλίου μεταξύ 0 και 1 οδηγεί σε μια καμπύλη. Σε ένα ιδανικό μοντέλο, η καμπύλη θα έφτανε σε όλη την αριστερή πλευρά και στην κορυφή, ώστε να καλύπτει όλη την περιοχή του διαγράμματος. Όσο μεγαλύτερη είναι η περιοχή κάτω από την καμπύλη (η οποία μπορεί να έχει οποιαδήποτε τιμή από 0 έως 1), τόσο καλύτερη είναι η απόδοση του μοντέλου. Επομένως, αυτή είναι η μετρική AUC που παρατίθεται μαζί με τις άλλες μετρικές παρακάτω.

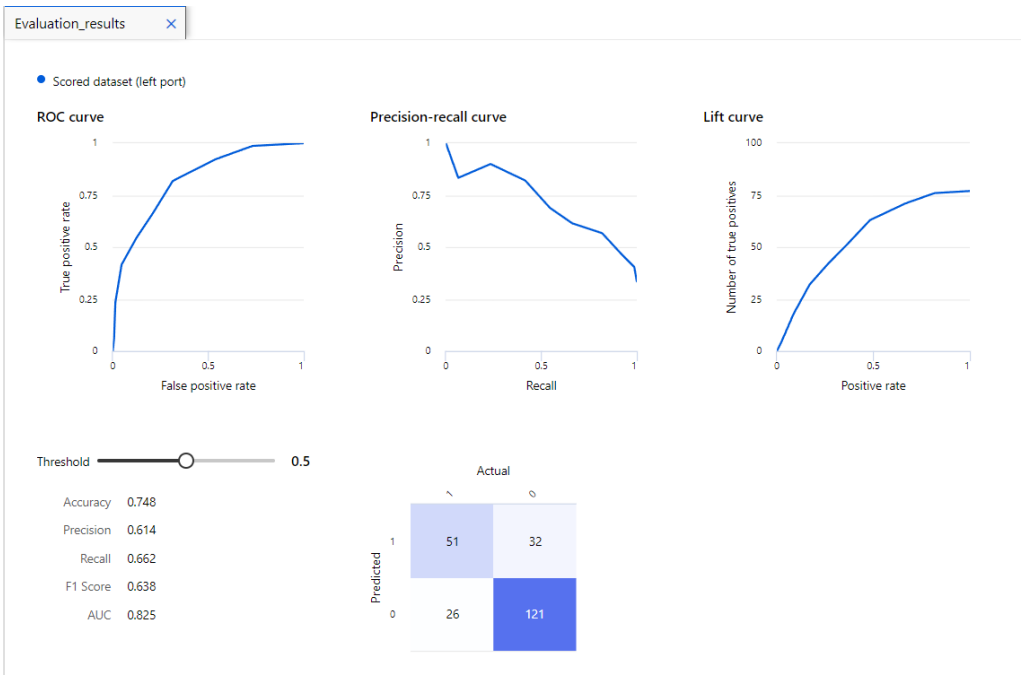
Για να πάρουμε μια ιδέα για το πώς αυτή η περιοχή αντιπροσωπεύει την απόδοση του μοντέλου, έχουμε για παράδειγμα μια ευθεία διαγώνια γραμμή από κάτω αριστερά προς τα πάνω δεξιά του διαγράμματος ROC. Αυτή αντιπροσωπεύει την αναμενόμενη απόδοση εάν απλά μαντεύαμε ή ρίχναμε ένα νόμισμα για κάθε ασθενή. Θα αναμέναμε να πετύχουμε περίπου τις μισές από αυτές σωστά και τις μισές λάθος, οπότε η περιοχή κάτω από τη διαγώνια γραμμή αντιπροσωπεύει μια AUC 0,5. Εάν η AUC για το μοντέλο μας είναι υψηλότερη από αυτό για ένα δυαδικό μοντέλο ταξινόμησης, τότε το μοντέλο αποδίδει καλύτερα από μια τυχαία εικασία.

11. Κλείνουμε το παράθυρο απεικόνισης αποτελεσμάτων Evaluate Model.

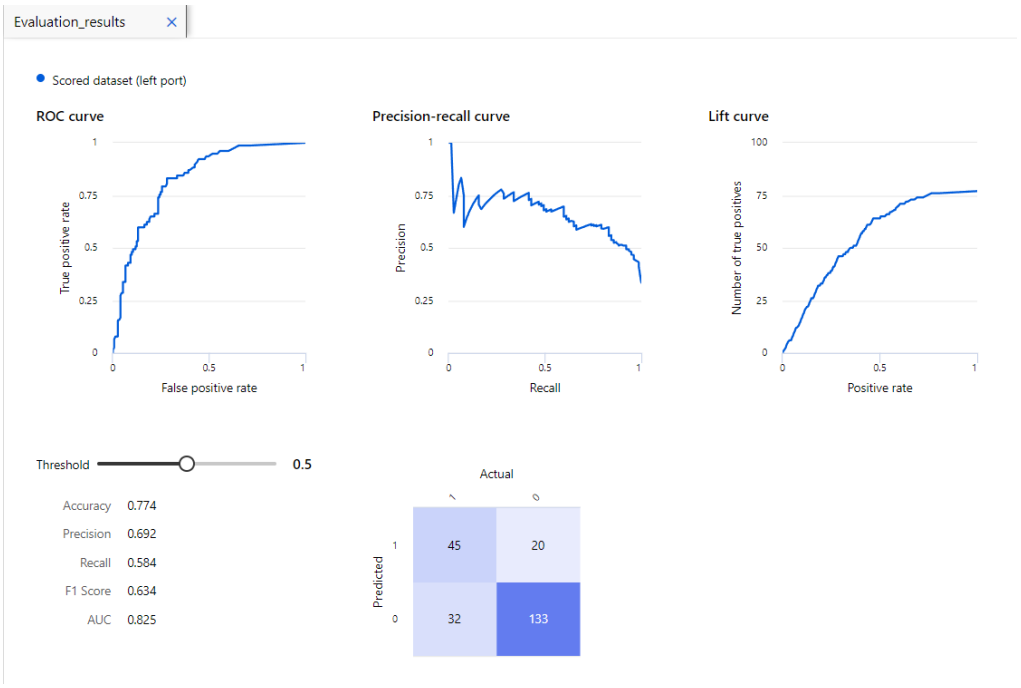
Η απόδοση αυτού του μοντέλου δεν είναι τόσο μεγάλη, εν μέρει επειδή πραγματοποιήσαμε μόνο ελάχιστη μηχανική και προ-επεξεργασία των χαρακτηριστικών του ασθενούς. Θα μπορούσαμε να δοκιμάσουμε έναν διαφορετικό αλγόριθμο ταξινόμησης, όπως τον Two-Class Decision Forest, και να συγκρίνουμε τα αποτελέσματα. Επίσης, μπορούμε να συνδέσουμε τις εξόδους της ενότητας Split Data σε πολλαπλές ενότητες Train Model και Score Model και μπορούμε να συνδέσουμε μια δεύτερη ενότητα Score Model στην ενότητα Evaluate Model για να δούμε μια σύγκριση δίπλα-δίπλα. Ο σκοπός της άσκησης είναι απλώς να σας εισαγάγει στην ταξινόμηση και στη διεπαφή του σχεδιαστή μηχανικής μάθησης Azure, όχι να εκπαιδεύσουμε ένα τέλειο μοντέλο.

Στη συνέχεια, θα δοκιμάσουμε διαφορετικούς αλγορίθμους στο ίδιο μοντέλο, ακολουθώντας ακριβώς την ίδια διαδικασία, μόνο που αυτή την φορά θα χρησιμοποιηθεί ένα άλλο μικρότερο dataset. Παρακάτω θα δούμε τα αποτελέσματα και την αξιολόγηση για κάθε αλγόριθμο.

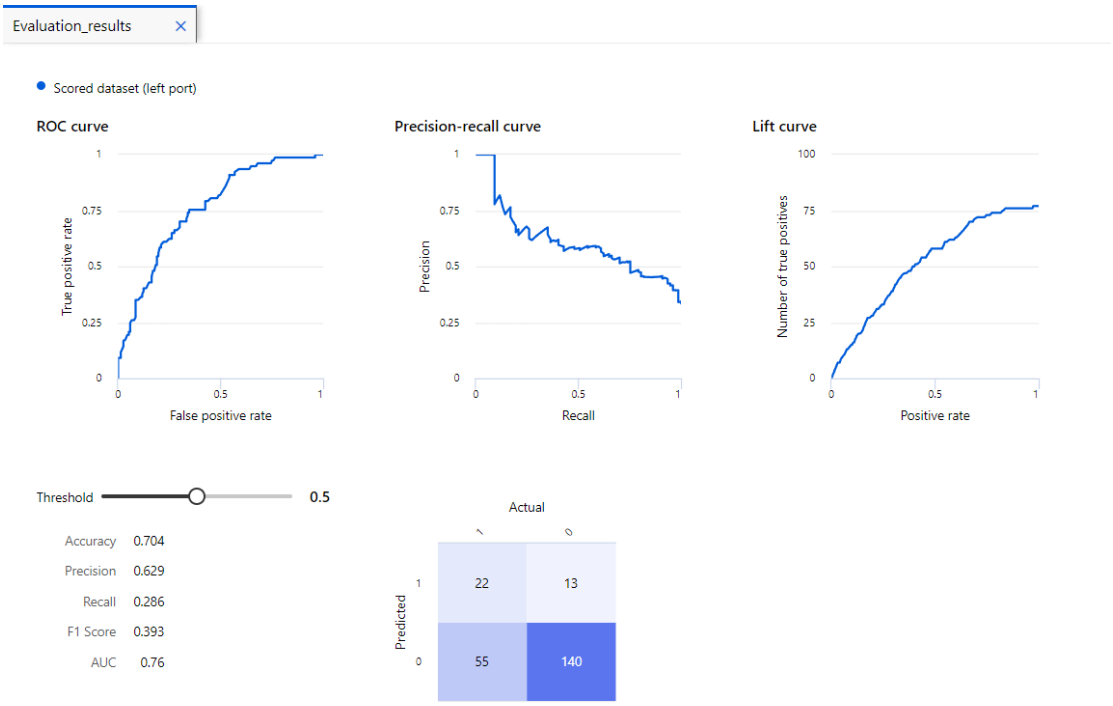
- Two-Class Decision Forest



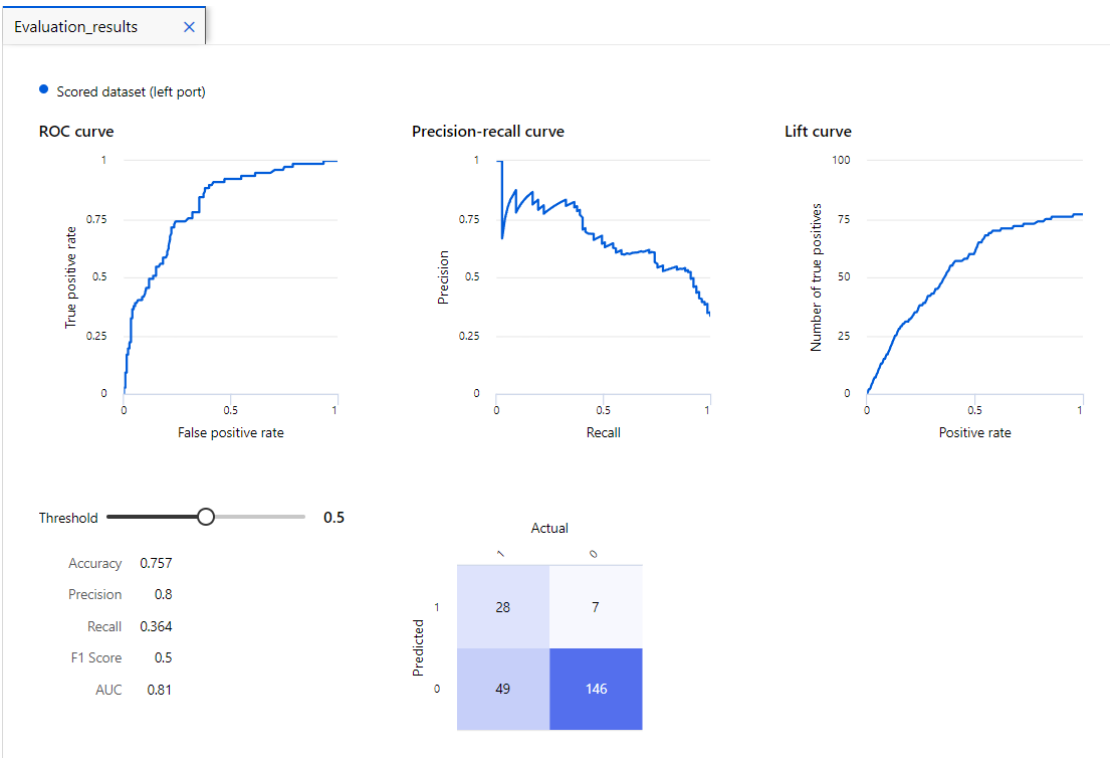
- Two-Class Boosted Decision Tree



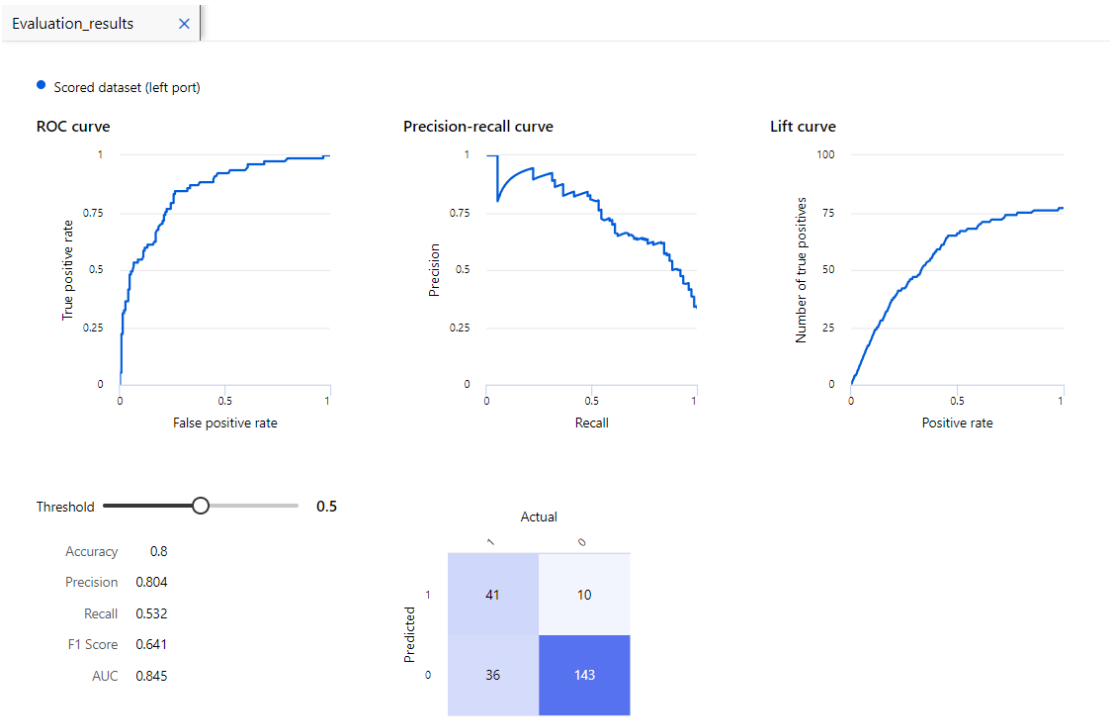
- Two-Class Support Vector Machine



- Two-Class Neural Network



- Two-Class Logistic Regression



Έπειτα, παρατηρούμε τις αποδόσεις των αλγορίθμων στον παρακάτω πίνακα:

Evaluation Results	Two-Class Decision Forest	Two-Class Boosted Decision Tree	Two-Class Support Vector Machine	Two-Class Neural Network	Two-Class Logistic Regression
Accuracy	0.748	0.774	0.704	0.757	0.8
Precision	0.614	0.692	0.629	0.8	0.804
Recall	0.662	0.584	0.286	0.364	0.532
F1 Score	0.638	0.634	0.393	0.5	0.641
AUC	0.825	0.825	0.76	0.81	0.845

Με βάση τον πίνακα, διαπιστώνουμε πως για το συγκεκριμένο σενάριο και dataset, ο αλγόριθμος με την μεγαλύτερη ακρίβεια (80%) είναι ο Two-Class Logistic Regression. Επομένως, επιλέγουμε αυτόν τον αλγόριθμο για το μοντέλο που δημιουργήσαμε, ώστε με οποιοδήποτε άλλο σύνολο δεδομένων να έχουμε τα βέλτιστα αποτελέσματα. Αυτό όμως δεν σημαίνει πως κάθε φορά θα είναι ο βέλτιστος αλγόριθμος. Ανάλογα με την φύση του προβλήματος εκτελούμε μια γκάμα αλγορίθμων και επιλέγουμε πάντα αυτόν με το μεγαλύτερο ποσοστό ακρίβειας. Δεν είμαστε σε θέση να γνωρίζουμε ποιος αλγόριθμος θα ανταπεξέλθει καλύτερα στο εκάστοτε πρόβλημα. Επομένως, η διαδικασία εύρεσης του κατάλληλου αλγορίθμου πραγματοποιείται με σκοπό να έχουμε τις πιο σωστές προβλέψεις, οι οποίες θα βρίσκονται όσο το δυνατόν πιο κοντά στα πραγματικά δεδομένα και γεγονότα.

5. Αξιολόγηση Πλατφόρμας

Η χρήση των εργαλείων μηχανικής μάθησης παρουσιάζει σήμερα μια σημαντική χρησιμότητα για επιχειρήσεις, είτε ιδιωτικές είτε δημόσιες. Η εφαρμογή τους μπορεί να αποτελέσει ένα σημαντικό πλεονέκτημα έναντι των ανταγωνιστών. Επιπλέον, η ακριβής πρόβλεψη των τάσεων μας παρέχει τη δυνατότητα να κατευθύνουμε τις ενέργειές μας προς τη σωστή κατεύθυνση και να μειώσουμε τυχόν ανεπιθύμητες επιπτώσεις. Το Microsoft Azure προσφέρει εργαλεία που ανιχνεύουν μοτίβα και, μέσω μοντέλων μηχανικής μάθησης, επιτρέπουν την πρόβλεψη μελλοντικών εξελίξεων. Η ευκολία χρήσης του Azure επιτρέπει την άνετη εισαγωγή δεδομένων από διάφορες πηγές, κάνοντας τη χρήση του από κάθε οργανισμό πιο προσβάσιμη. Η χρήση του δίνει ισχυρό πλεονέκτημα στις επιχειρήσεις λόγω του μειωμένου κόστους των υπηρεσιών του έναντι άλλων συστημάτων. Επιπλέον, παρέχει δυνατότητες επεξεργασίας δεδομένων με διαφορετικούς αλγορίθμους, ανάλογα με τις απαιτήσεις των δεδομένων και την δυνατότητα προβλέψεων μελλοντικών τιμών και συμπεριφορών. Μέσω του Azure είναι εφικτή η σύγκριση της αποτελεσματικότητας των αλγορίθμων στα ίδια δεδομένα και η επιλογή του κατάλληλου ανάλογα με τις ανάγκες μας. Παρόλα αυτά, η λανθασμένη ή βεβιασμένη χρήση του Azure μπορεί να έχει αρνητικές συνέπειες για τον οργανισμό. Αυτό συμβαίνει για την παρουσιάζει μια πολύπλοκη διαδικασία για μη εξοικειωμένους χρήστες στην ανάπτυξη των υπηρεσιών επειδή είναι στο cloud. Επίσης, είναι σημαντικό οι χρήστες να διαθέτουν ένα καλό σύνολο δεδομένων, έτσι ώστε η εκπαίδευση του μοντέλου να είναι αποτελεσματική και να παρέχει αξιόπιστα αποτελέσματα. Η έλλειψη ενός ικανοποιητικού συνόλου δεδομένων μπορεί να οδηγήσει σε εσφαλμένες προβλέψεις και να αποτρέψει τον οργανισμό από τη λήψη σωστών αποφάσεων. Συνεπώς, η συνετή χρήση των εργαλείων του Microsoft Azure είναι κρίσιμη για την επίτευξη θετικών αποτελεσμάτων.

6. Συμπεράσματα

Στόχος της εργασίας ήταν να εξερευνήσουμε και να αποκτήσουμε εκτενή κατανόηση του ευρύτερου πεδίου της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης. Αναλύσαμε ενδελεχώς αυτούς τους δύο συνδεδεμένους τομείς. Μέσω της πλατφόρμας του Microsoft Azure και της υλοποίησης σεναρίων, καταλήξαμε σε σημαντικά ευρήματα τόσο για τις εφαρμογές αυτές όσο και για το ευρύτερο πεδίο που εξετάσαμε.

Η τεχνητή νοημοσύνη και η μηχανική μάθηση αποτελούν δύο τομείς στους οποίους κάθε οργανισμός που επιθυμεί να παρακολουθεί και να προβλέπει τις εξελίξεις πρέπει να επικεντρωθεί. Καθώς προχωράμε στην τρίτη δεκαετία του 21ου αιώνα, οι πηγές δεδομένων αυξάνονται διαρκώς. Αυτές αποτελούν το πιο προσιτό εργαλείο κάθε οργανισμού, το οποίο πρέπει να εκμεταλλευτεί προκειμένου να αξιοποιήσει όσο το δυνατόν καλύτερα τις ευκαιρίες, χωρίς να περιορίζεται μόνο στο οικονομικό κέρδος. Αξίζει να σημειωθεί ότι, όταν αναφερόμαστε σε δημόσιους οργανισμούς, ο στόχος τους δεν είναι πάντα ο οικονομικός κέρδος, αλλά η βελτίωση της εξυπηρέτησης του κοινού ή των πολιτών. Το πιο δύσκολο και χρονοβόρο μέρος της διαδικασίας αυτής είναι η απόκτηση των κατάλληλων δεδομένων και η σωστή τους επεξεργασία, προκειμένου να προβλέπουμε τις μελλοντικές συμπεριφορές με τη μέγιστη δυνατή ακρίβεια. Αντλώντας και αποθηκεύοντας αυτά τα δεδομένα απαιτείται ένας δαπανηρός και χρονοβόρος φόρτος εργασίας, παρόλο που πρόσφατα η ανάπτυξη του αποθηκευτικού χώρου έχει αρχίσει να αντιμετωπίζει λιγότερα προβλήματα. Τα λάθη μπορούν να εμφανιστούν σε διάφορα επίπεδα, από την απόκτηση και τη δομή των δεδομένων έως τη δημιουργία κατάλληλων συσχετίσεων μεταξύ τους, καθώς και στην αναγνώριση των προβλέψεων. Καταλήγουμε στο συμπέρασμα ότι η ύπαρξή εργαλείων, όπως το Azure, σε συνδυασμό με μια σωστή και προσεκτική χρήση, μπορεί να οδηγήσει τον οργανισμό σε σημαντικά οφέλη. Αυτά τα οφέλη μπορούν να φτάσουν ακόμη και σε επίπεδο ανταγωνιστικού πλεονεκτήματος έναντι άλλων επιχειρήσεων και να καταστήσουν τον οργανισμό μαγνήτη για καταναλωτές και εργαζομένους. Επιπλέον, η χρήση τέτοιων εργαλείων δημιουργεί ένα φιλικότερο τεχνολογικά εργασιακό περιβάλλον.

Συνολικά, η αξιοποίηση της τεχνητής νοημοσύνης και της μηχανικής μάθησης αποτελεί κρίσιμο παράγοντα για την επιτυχία και την ανταγωνιστικότητα των επιχειρήσεων σήμερα. Με την κατάλληλη διαχείριση και εφαρμογή των εργαλείων μηχανικής μάθησης, μπορούν να προκύψουν σημαντικά οφέλη τόσο σε οικονομικό επίπεδο όσο και στην ποιότητα της εξυπηρέτησης των πελατών ή των πολιτών. Είναι, λοιπόν, σημαντικό να επιδείξουν οι επιχειρήσεις προσεκτικότητα και σύνεση στη χρήση αυτών των εργαλείων, προκειμένου να αξιοποιήσουν πλήρως τα πλεονεκτήματά τους και να αποφύγουν πιθανούς κινδύνους, δυσκολίες και παραπλανητικά αποτελέσματα.

Βιβλιογραφία

Armbrust, Michael, et al. "Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics." Proceedings of CIDR. Vol. 8. 2021.

Barga, Roger, et al. Predictive analytics with Microsoft Azure machine learning. Berkely, CA: Apress, 2015.

Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006.

Davenport, Thomas H., and Jeanne G. Harris. "Competing on analytics: the new science of Winning." Language 15.217p (2007): 24cm.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

Hardesty, Larry. "Explained: neural networks." MIT News 14 (2017).

Hendrickson, Scott, et al. "Serverless computation with {OpenLambda}." 8th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 16). 2016.

Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.

Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου (2011). «Τεχνητή Νοημοσύνη», Αθήνα: Εκδόσεις Πανεπιστημίου Μακεδονίας.

INMON, WH. "Building the Warehouse, 2nd." (1996): E23.

Khine, Pwint Phyu, and Zhao Shun Wang. "Data lake: a new ideology in big data era." ITM web of conferences. Vol. 17. EDP Sciences, 2018.

Kimball, Ralph, and Margy Ross. The data warehouse toolkit: the definitive guide to dimensional modeling. John Wiley & Sons, 2013.

Kimball, Ralph. The data warehouse toolkit: practical techniques for building dimensional data warehouses. John Wiley & Sons, Inc., 1996.

Kubat, Miroslav. An introduction to machine learning. Springer, 2017.

Mazumdar, Dipankar, Jason Hughes, and J. B. Onofre. "The Data Lakehouse: Data Warehousing and More." arXiv preprint arXiv:2310.08697 (2023).

Miloslavskaya, Natalia, and Alexander Tolstoy. "Application of big data, fast data, and data lake concepts to information security issues." 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW). IEEE, 2016.

Mitchell, Tom M. "Does machine learning really work?." AI magazine 18.3 (1997): 11-11.

Pannu, Avneet. "Artificial intelligence and its application in different areas." Artificial Intelligence 4.10 (2015): 79-84.

Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. An introduction to machine learning. Springer, 2019.

Russell, S. J., and Norvig, P. Artificial intelligence - a modern approach (4th edition). London, 2021.

Timothy King in Best Practices, Data Warehouse vs. Data Lake; What's the Difference? June 9, 2016, Retrieved Sep 10, 2017: https://solutionsreview.com/datamanagement/data-warehouse-vs-data-lake-whats-the-difference/.

Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. "Statistical learning theory." (1998): 1780.

Yessad, Lamia, and Aissa Labiod. "Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault." 2016 International Conference on System Reliability and Science (ICSRS). IEEE, 2016.

Γεωργούλη, Α. (2015). Τεχνητή νοημοσύνη [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://dx.doi.org/10.57713/kallipos-666

Ιστοσελίδες

[1] Amazon Athena. http://aws.amazon.com/athena/.

[2] Amazon Glue. https://aws.amazon.com/glue/.

[3] Apache Hudi. https://hudi.apache.org/

[4] Apache Iceberg: The open table format for analytic datasets https://iceberg.apache.org/

[5] Dedicated SQL Pool. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-what-is?context=%2Fazure%2Fsynapse-analytics%2Fcontext%2Fcontext

[6] Delta Lake. https://delta.io/

[7] Google BigQuery. https://cloud.google.com/bigquery/.

[8] OpenWhisk. https://developer.ibm.com/openwhisk/.

[9] Serverless SQL Pool. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-serverless-sql-pool

[10] Serverless SQL Pool. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/on-demand-workspace-overview

[11] Azure Synapse Analytics https://learn.microsoft.com/en-us/azure/synapse-analytics/overview-what-is