



راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده

شرکت رایانش سریع هزاره ایرانیان



شناسنامه سند

عنوان سند	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده
شناسه ی سند	

تأیید و تصویب سند

نام و نام خانوادگی	سمت	تاریخ	امضاء
تهیه کننده	دکتر امین نظارات	۳۰ دیماه ۱۳۹۹	
	مهدیه معمارزاده	۲۰ دیماه ۱۳۹۹	
تصویب کننده	دکتر امین نظارات	۱۵ بهمن ۱۳۹۹	


سوابق ویرایش

شماره ویرایش	تاریخ انتشار	خلاصه تغییرات
۱,۰,۰	۲۰ دیماه ۱۳۹۹	آماده سازی اولیه سند
۲,۰,۰	۱۵ بهمن ۱۳۹۹	افزودن بخشهای امکان سنجی

تماس با ما



آدرس پستی	یزد، خیابان آیت الله کاشانی، کوچه ۲۹، شرکت رایانش سریع هزاره ایرانیان، کدپستی ۸۹۱۵۶۶۵۴۸۵
تلفن تماس	۰۳۵۳۶۲۳۲۱۷۶
وبسایت	www.astek.ir
پست الکترونیک	info@astek.ir



	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

این مستند به منظور ارائه یک راهنمای کامل برای مستند کنترل پروژه با شماره سند BD.PRP.PM.012.00 می باشد. در این مستند تمامی فعالیتهای یک پروژه علم داده که مقرر است اجرا شود به تشریح بیان شده است و وظایف تیمهای مختلف در آن شرح داده شده است.

Task Name	Description
Data Science Project	این فایل یک الگوی عمومی برای استفاده در پروژه های علم داده است.
Project Kickoff	این چرخه حیات برای پروژه های علوم داده طراحی شده است که قرار است به عنوان بخشی از کاربردهای هوش مصنوعی در سازمان عملیاتی شوند. این برنامه ها از مدل های یادگیری ماشین یا هوش مصنوعی برای تجزیه و تحلیل پیش بینی استفاده می کنند. پروژه های علوم داده اکتشافی و پروژه های تجزیه و تحلیل موقت یا غیرفعال نیز می توانند از استفاده از این فرآیند بهره مند شوند، اما در چنین مواردی برخی از مراحل شرح داده شده ممکن است لازم نباشد.
Business Understanding	<p>اهداف</p> <ul style="list-style-type: none"> - متغیرهای کلیدی مشخص شده ای هستند که به عنوان اهداف مدل عمل می کنند و معیارهای مربوط به آنها برای موفقیت پروژه تعیین می شود. - منابع داده ای شناسایی می شوند که کسب و کار به آنها دسترسی دارد یا باید آنها را بدست آورد. <p>چگونه انجامش دهیم</p> <p>در این مرحله دو وظیفه اصلی وجود دارد:</p> <ul style="list-style-type: none"> - اهداف را تعریف کنید: برای درک و شناسایی مشکلات تجاری با مشتری و سایر ذینفعان کار کنید. - فرمول هایی را تعریف کنید که اهداف تجاری را مشخص می کند و تکنیک های علم داده می توانند آنها را هدف قرار دهند. - منابع داده را شناسایی کنید: داده های مربوطه را پیدا کنید که به شما کمک می کند به سوالات تعیین کننده اهداف پروژه پاسخ دهید.
Ideation	<p>برخی از مهمترین کارها در چرخه کلی حیات پروژه های داده قبل از نوشتن یک خط کد اتفاق می افتد. اگر به خوبی انجام شود، مرحله ایده پردازی با رفت و برگشت بین طرف های ذینفع، به طرز چشمگیری یک پروژه را کم ریسک می کند. اینجاست که هدف تجاری مشخص می شود، معیارهای موفقیت تعیین می شود، پروژه های شاخص قبلی بررسی می شوند و محاسبات اولیه بازگشت سرمایه انجام می شود.</p> <p>ایده زمانی است که امکان سنجی ارزیابی شود، هر دو از نظر</p> <p>"آیا داده ها وجود دارند؟"</p>

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	 
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

و

"آیا می توانیم واقعاً نحوه کار فرآیند کسب و کار را تغییر دهیم؟"

همچنین جایی است که اولویت بندی نسبت به سایر پروژه های بالقوه اتفاق می افتد. در زیر برخی از بهترین اقدامات مشاهده شده وجود دارد که ریشه بسیاری از مشکلاتی است که گفته شد.


اول مسئله، نه اول داده

بسیاری از سازمانها به جای ایجاد درک عمیق از روند کسب و کار موجود و سپس مشخص کردن نقطه تصمیم گیری که می تواند تقویت یا خودکار شود، با داده ها شروع می کنند و به دنبال چیزی "جذاب" می گردند. پیدایش یک پروژه نه تنها از کسب و کار لازم است بلکه باید متناسب با یک مشکل خاص تجاری باشد.



فرآیندهای موجود را نگاشت دهید

سازمانهای پیشرو فرایندهای تجاری موجود را ابتدا شناسایی کرده و سپس نقاط دقیقی را که علم داده می تواند تأثیر تجاری داشته باشد، بر روی آن محک می زنند. با این کار تضمین می شود که آنها برای هدف قرار دادن بخشهای تأثیرگذارتر چرخه، فرصت را از دست نمی دهند. همچنین این موضوع اطمینان می دهد که آنها در طول چرخه زندگی با زبان ذینفعان خود صحبت می کنند تا مدیریت تغییرات را به حداقل برسانند.

Define Objectives	<p>برای درک و شناسایی مشکلات تجاری با مشتری و سایر ذینفعان کار کنید. سؤالاتی را فرموله کنید که اهداف تجاری را مشخص می کند و تکنیک های علم داده می توانند آنها را هدف قرار دهند.</p>
Identify the key business variables	<p>هدف اصلی این مرحله شناسایی متغیرهای اصلی کسب و کار است که از منظر تجزیه و تحلیل نیاز به پیش بینی دارد. از این متغیرها به عنوان اهداف مدل یاد می شود و از معیارهای مرتبط با آنها برای تعیین موفقیت پروژه استفاده می شود. دو نمونه از این اهداف پیش بینی فروش یا پیش بینی سفارش بعدی است.</p>
Define Project Goals	<p>با طرح سؤالات "شارپ" مرتبط و مشخص و بدون ابهام، اهداف پروژه را مشخص کنید. علم داده فرآیند استفاده از نامها و اعداد برای پاسخگویی به چنین سؤالات است.</p> <p>علم داده / یادگیری ماشینی به طور معمول برای پاسخ به پنج نوع سال استفاده می شود:</p> <ul style="list-style-type: none"> •How much or how many? (regression) •Which category? (classification) •Which group? (clustering) •Is this weird? (anomaly detection)

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

	<p>•Which option should be taken? (recommendation)</p> <p>مشخص کنید کدام یک از این سوال ها را می پرسید و پاسخ دادن به آن چگونه اهداف تجاری سازمان را محقق می کند.</p>
Define project team	<p>تیم پروژه را با تعیین نقش ها و مسئولیت های اعضای آن مشخص کنید. یک طرح مایل استون سطح بالا تهیه کنید که برای کشف اطلاعات بیشتر، آن را تکرار کنید.</p>
Define success metrics	<p>معیارهای موفقیت را تعریف کنید.</p> <p>به عنوان مثال، در پایان این پروژه ۳ ماهه به درصد دقت ایکس٪ در پیش بینی ریزش مشتریان خواهیم رسید، بنابراین می توانیم پیشنهادات بهتری را به مشتریان جهت جلوگیری از ریزش ارائه کنیم.</p> <p>این معیارها باید خصوصیات زیر را داشته باشند:</p> <p>SMART:</p> <ul style="list-style-type: none"> • Specific • Measurable • Achievable • Relevant • Time- bound
Identify Data Sources	<p>منابع داده ای را که شامل نمونه های شناخته شده ای از پاسخ به سوالات شارپ شما هستند، شناسایی کنید. به دنبال داده های زیر باشید:</p> <p>– داده هایی که به سوال مربوط می شوند. آیا معیارهایی از هدف و ویژگی های (فیچرها) مرتبط با هدف را داریم؟</p> <p>– داده هایی که معیارهای دقیق هدف نهایی ما را شامل شده باشند و فیچرهای مورد نظر را هم داشته باشند.</p> <p>بعید نیست، به عنوان مثال، کشف اینکه سیستم های موجود برای حل مشکل و دستیابی به اهداف پروژه نیاز به جمع آوری و ثبت انواع داده های اضافی باشد. در این حالت، ممکن است بخواهید به دنبال منابع داده خارجی باشید یا سیستم های خود را برای جمع آوری داده های جدید به روز کنید.</p>
Determine project feasibility	<p>برخی از سوالات مفید را برای تعیین امکان سنجی یک پروژه باید پرسید:</p> <p>هزینه کسب اطلاعات</p> <p>– دستیابی به داده ها چقدر سخت است؟</p> <p>– برچسب گذاری داده ها چقدر گران است؟</p> <p>– چه مقدار داده مورد نیاز خواهد بود؟</p>

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	 
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

در دسترس بودن کارهای منتشر شده خوب درباره مشکلات مشابه

- آیا مشکل به تمرین خلاصه شده است؟
- آیا در مورد مسئله پژوهشهای کافی وجود دارد؟

منابع محاسباتی هم برای آموزش و هم برای ارزیابی در دسترس است

- آیا این مدل در محیط با منابع محدود هم استفاده می شود؟

Data Sources Documented

این بخش به این موضوع می پردازد که آیا منابع داده ای اولیه به خوبی مستند شده اند و گزارش دیتا دیفینیشن برای آنها تهیه شده است؟

Data Dictionaries Delivered

این سند شرح داده هایی است که توسط مشتری ارائه می شود. این توصیفات شامل اطلاعات مربوط به طرحواره (انواع داده ها، اطلاعات در مورد قوانین اعتبار سنجی در صورت وجود) و نمودارهای رابطه موجودیت در صورت موجود بودن است.

Charter Document Delivered

این یک سند زنده و جاری است که با یافته های جدید و تغییر الزامات تجاری در طول پروژه به روز می شود. نکته اصلی، تکرار این سند و افزودن جزئیات بیشتر است، همانطور که در روند کشف مفاهیم پیشرفت می کنید مشتری و سایر ذینفعان را درگیر ایجاد تغییرات کنید و دلایل تغییرات را به روشنی بیان کنید.


Setting up a ML codebase

یک پایگاه کد یادگیری ماشین به خوبی سازمان یافته باید پردازش داده ها، تعریف مدل، آموزش مدل و مدیریت آزمایش را مدوله کند.
این پایگاه کد می تواند به صورت زیر باشد:


```

├── README.md    <- You are here
├── config        <- Directory for yaml configuration files for model training, scoring, etc
├── logging/      <- Configuration of python loggers
├── data          <- Folder that contains data used or generated. Only the external/ and sample/ subdirectories are tracked by git.
│   ├── archive/  <- Place to put archive data is no longer usable. Not synced with git.
│   ├── external/ <- External data sources, will be synced with git
│   └── sample/   <- Sample data used for code development and testing, will be synced with git
├── docs          <- A default place for project's documents.
└── figures       <- Generated graphics and figures to be used in reporting.


```

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	 <p>شرکت رایانش سریع هزاره ایرانیان</p>
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir


	<ul style="list-style-type: none"> models <- Trained model objects (TMOs), model predictions, and/or model summaries archive <- No longer current models. This directory is included in the .gitignore and is not tracked by git notebooks <ul style="list-style-type: none"> develop <- Current notebooks being used in development. deliver <- Notebooks shared with others. archive <- Develop notebooks no longer being used. template.ipynb <- Template notebook for analysis with useful imports and helper functions. src <- Source data for the project <ul style="list-style-type: none"> archive/ <- No longer current scripts. helpers/ <- Helper scripts used in main src files sql/ <- SQL source code ingest_data.py <- Script for ingesting data from different sources generate_features.py <- Script for cleaning and transforming data and generating features used for use in training and scoring. train_model.py <- Script for training machine learning model(s) score_model.py <- Script for scoring new predictions using a trained model. postprocess.py <- Script for post processing predictions and model results evaluate_model.py <- Script for evaluating model performance test <- Files necessary for running model tests <ul style="list-style-type: none"> true <- Directory containing sources of truth for what results produced in each test should look like test <- Directory where artifacts and results of tests are saved to be compared to the sources of truth. Only gitkeep in this directory should be synced to Git test.py <- Runs the tests defined in test_config.yml and then compares the produced artifacts/results with those defined as expected in the true/ directory test_config.yml <- Configures the set of tests for comparing artifacts and results. Currently does not include unit testing or other traditional software testing run.py <- Simplifies the execution of one or more of the src scripts requirements.txt <- Python package dependencies
Data Acquisition and Understanding	<p>اهداف:</p> <p>یک مجموعه داده تمیز و با کیفیت بالا که روابط آن با متغیرهای هدف قابل درک است و در محیط تجزیه و تحلیل مناسب و آماده مدل سازی قرار دارند.</p> <p>یک معماری راه حل از پایپ لاین داده برای تازه سازی و امتیاز دهی به طور منظم داده شده است.</p>
Ingest the data	<p>فرایندهایی را تنظیم کنید تا داده ها را از مکان های منبع اصلی به مکان های هدف که در آن عملیات تجزیه و تحلیل مانند آموزش و پیش بینی ها اجرا می شود، انتقال دهید.</p>

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir



Explore the data	<p>قبل از اینکه مدل های خود را آموزش دهید، باید درک صحیحی از داده ها داشته باشید. مجموعه داده های دنیای واقعی غالباً پر از نویز هستند یا مقادیر از دست رفته دارند و یا انبوهی از اختلافات دیگر را دارند. از خلاصه سازی و تجسم داده می توان برای ارزیابی کیفیت داده ها و ارائه اطلاعات مورد نیاز برای پردازش داده ها قبل از آماده شدن برای مدل سازی، استفاده کرد. این روند اغلب تکراری است.</p> <p>هنگامی که از کیفیت داده های پاک شده رضایت پیدا کردید، گام بعدی درک بهتر الگوهایی است که در داده ها وجود دارد و به شما کمک می کند تا یک مدل پیش بینی مناسب برای هدف خود انتخاب و توسعه دهید. به دنبال شواهدی برای ارتباط خوب داده ها با هدف و اینکه آیا داده های کافی برای پیشبرد مراحل بعدی مدل سازی وجود دارد، باشید. باز هم، این روند اغلب تکراری است. برای تقویت مجموعه داده ای که در مرحله قبل مشخص شده است، ممکن است لازم باشد منابع داده جدیدی با داده های دقیق تر یا مرتبط تر پیدا کنید.</p>
Define ground truth (create labeling documentation)	<p>برای کسب نتایج قابل قبول از یک مدل یادگیری ماشینی وجود داده های مطمئن و کافی می تواند تاثیر زیادی در پذیرش نتایج توسط مشتری داشته باشد. داده های دارای برچسب از جمله ضروریات برای دستیابی به دقت کافی است. در صورت عدم وجود برچسب در کل پروژه، حصول نتیجه قابل قبول با ابهام روبروست. در این فعالیت لازم است که تیم مجری نسبت به بررسی وضعیت وجود/عدم وجود یا کفایت/عدم کفایت برچسبهای داده های مورد نیاز اقدام نمایند</p> <p>در صورت عدم وجود برچسبهای کافی برای داده ها می تواند از یادگیری فعال استفاده کرد.</p> <p>-Active Learning</p> <p>هنگامی که مقدار زیادی داده بدون برچسب دارید مفید است، باید تصمیم بگیرید که چه داده هایی را باید برچسب گذاری کنید. برچسب گذاری داده ها می تواند گران تمام شوند، بنابراین ما می خواهیم زمان صرف شده برای این کار را محدود کنیم.</p> <p>قبل از اجرای هر الگوریتم پیشرفته یادگیری ماشینی، دانشمندان داده یا تیم های داده اغلب به داده های دارای برچسب نیاز دارند. نیازی به گفتن نیست که این فرایند جمع آوری بسیار خسته کننده و زمان بر است و در اینجا یادگیری فعال به کمک می آید. یادگیری فعال الگوریتم هایی را برای اتوماسیون برچسب گذاری داده ها اعمال می کند.</p> <p>احتمالاً، یادگیری فعال می تواند بسیار فوری و مفید عمل کند، خصوصاً در مواردی که داده های بدون برچسب زیادی وجود دارد که بسیار گران قیمت هستند و یا برای برچسب زدن با دست بسیار وقت گیر باشند. با این حال، هنوز کاملاً مشخص نیست که یادگیری فعال چه زمانی کار می کند، بر روی کدام داده ها و کدام تکنیک ها بهترین عملکرد را دارند. با افزایش تعداد داده ها در همه دامنه ها، برچسب گذاری داده ها به یک گلوگاه تبدیل می شود. مشاغلی که می توانند به این مسئله بپردازند تا سریعاً داده های قابل استفاده برای یادگیری تحت نظارت داشته باشند، قادر خواهند بود سریعترین مقیاس را داشته باشند. یادگیری فعال می تواند برای بسیاری از صنایع مفید باشد، به ویژه برای آنهایی که دارای داده اینترنت اشیا هستند (به عنوان مثال، تولید) و برای هر پروژه ای که شامل داده های غیر ساختاری مانند تصاویر، فیلم ها و غیره باشد.</p>

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir


	<p>رویکرد کلی:</p> <p>۱- با شروع یک مجموعه داده بدون برچسب، با به دست آوردن برچسب برای زیر مجموعه کوچکی از نمونه ها، یک مجموعه داده "دانه" ایجاد کنید.</p> <p>۲- مدل اولیه را بر روی مجموعه داده های دانه آموزش دهید</p> <p>۳- برچسب مشاهدات بدون برچسب باقی مانده را پیش بینی کنید</p> <p>۴- از عدم اطمینان پیش بینی های مدل برای اولویت بندی برچسب گذاری مشاهدات باقی مانده استفاده کنید</p>
Set up data pipeline	<p>علاوه بر واکنشی اولیه و تمیز کردن داده ها، شما معمولاً باید فرایندی را برای امتیاز دهی داده های جدید یا تازه سازی داده ها به طور منظم به عنوان بخشی از یک فرایند یادگیری مداوم تنظیم کنید. این کار را می توان با راه اندازی پایپ لاین داده یا گردش کار انجام داد.</p> <p>در این مرحله یک معماری راه حل از پایپ لاین داده ایجاد شده است. پایپ لاین نیز به موازات اجرای پروژه توسعه می یابد. پایپ لاین ممکن است بسته به نیازهای تجاری شما و محدودیت های سیستم های موجود سازمان که این راه حل در آن ادغام شده است، به صورت دسته ای یا جریان/در زمان واقعی یا ترکیبی باشد.</p>
Deliver Data Quality Report	این گزارش شامل خلاصه داده ها، روابط بین هر ویژگی و هدف، رتبه بندی متغیر و غیره است.
Deliver a Solution Architecture	این بخش می تواند نمودار یا توصیفی از پایپ لاین داده پروژه باشد که برای اجرای امتیازدهی یا پیش بینی داده های جدید پس از ساختن یک مدل استفاده شود. همچنین می تواند شامل پایپ لاین برای آموزش مجدد مدل طراحی شده بر اساس داده های جدید باشد.
Checkpoint Decision	قبل از شروع مهندسی کامل ویژگی ها و ساخت مدل، می توانید پروژه را مجدداً ارزیابی کنید و تعیین کنید که آیا مقدار داده مورد انتظار برای ادامه کار کافی است یا خیر. به عنوان مثال شما ممکن است برای ادامه کار آماده باشید، نیاز به جمع آوری داده های بیشتر داشته باشید یا پروژه را رها کنید زیرا داده ها برای پاسخ دادن به سوال وجود ندارند.
Pre- Modeling	با توجه به تنوع مدلها برای حل یک مسئله، انتخاب مدل مناسب می تواند نتایج قابل قبول تری را به دنبال داشته باشد. هدف از این مرحله یک ارزیابی اولیه بر روی انواع مدلهای تجربه شده در دنیا در صنایع مشابه و کمک به مرحله انتخاب مدل می باشد
Models Exploration	بررسی تجربه صنایع و سازمانهای مشابه در حل مسئله جاری و مطالعه و بررسی مقالات و پژوهش های انجام شده در زمینه مربوطه به منظور دستیابی به آخرین تجربیات و نتایج تحقیقاتی که می تواند در پروژه جاری مثر ثمر باشد.

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

	<p>خروجی این فعالیت گزارشی از اثبات نظریه است که می تواند در انتخاب مسیر پیش رو کمک کننده باشد</p> <p>خط مشی اصلی را برای مشکل خود تعیین کنید. خطوط پایه هم برای ایجاد یک حد پایین تر از عملکرد مورد انتظار (پایه مدل ساده) و هم برای ایجاد یک سطح عملکرد هدف (پایه انسانی) مفید است.</p> <p>خطوط ساده شامل مدلهای یادگیری موجود در سایکیت لرن (یعنی رگرسیون لجستیک با پارامترهای پیش فرض) یا حتی الگوریتمهای اکتشافی ساده (همیشه کلاس اکثریت را پیش بینی می کنند) است.</p> <p>پیچیدگی را ساده شروع کنید و به تدریج افزایش دهید. این موضوع به طور معمول شامل استفاده از یک مدل ساده است، اما همچنین می تواند شروع با نسخه ساده تری از کار شما باشد.</p> <p>برای یافتن و بررسی سوابق کارهای دیگران می توان اقدامات زیر را انجام داد</p> <p>رسی ادبیات موضوع:</p> <p>مقالاتی را جستجو کنید که معماری مدل را برای مشکلات مشابه توصیف می کنند و با سایر دانشمندان داده صحبت کنید تا ببینید کدام رویکردها در عمل بیشترین موفقیت را داشته اند. رویکردی از میان پیشنهادات را تعیین کنید و از آن به عنوان یک مدل پایه (آموزش داده شده روی مجموعه داده خود) استفاده کنید.</p> <p>یک نتیجه شناخته شده را تولید کنید:</p> <p>اگر از مدلی استفاده می کنید که به خوبی مطالعه شده است، اطمینان حاصل کنید که عملکرد مدل شما در مجموعه داده ای که معمولاً مورد استفاده قرار می گیرد با آنچه در منابع گزارش شده مطابقت دارد.</p> <p>- درک کنید که چگونه عملکرد مدل با داده های بیشتر مقیاس پذیر خواهد شد:</p> <p>عملکرد مدل را به عنوان تابعی از افزایش اندازه مجموعه داده برای مدل های پایه ای که کاوش کرده اید رسم کنید. با افزایش مقدار داده مورد استفاده برای آموزش، مشاهده کنید که عملکرد هر مدل چگونه مقیاس می گیرد.</p>
Run Proof of Concept	<p>به منظور ایجاد اطمینان از عملکرد و نتایج پژوهشهای سایرین و اطمینان از امکان اجرای مدلهای مختلف کاندید شده بر روی داده های پروژه جاری، در این فعالیت اقدام به پیاده سازی اولیه یکی از مدلهای کاندید با دقت حداقلی می شود.</p>
Refine Metrics	<p>به منظور حصول اطمینان از تطابق متریک های تعریف شده برای پروژه با نتایج اولیه بدست آمده، در این فعالیت لازم است که مقادیر هدف گذاری شده برای متریک ها بازبینی شده و بر اساس واقعیت موجود در داده ها بازنگری شوند.</p> <p>این هدف گذاری برای مدلهای می تواند به صورت مثالهای زیر باشد:</p> <p>برای دقت بیشتر بهینه کنید</p>

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	 
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

	<p>تأخیر پیش بینی زیر ۱۰ میلی ثانیه مدل به بیش از ۲۵۶ گیگابایت حافظه نیاز ندارد ۹۰ درصد پوشش (برای اینکه یک پیش بینی را معتبر بشمریم حد آستانه ای برای اطمینان مدل تعیین می کنیم)</p>
Refine Proposal	<p>پس از اطمینان از امکان پذیر بودن اجرای پروژه از منظر میزان دقت مدل و کفایت داده ها، تیم می تواند اقدام به نهایی کردن شرح خدمات و جزئیات پروژه نماید.</p>
Modeling	<p>اهداف:</p> <ul style="list-style-type: none"> - ویژگی های بهینه داده ها برای مدل یادگیری ماشینی. - یک مدل یادگیری ماشینی آموزنده که هدف را به طور دقیق پیش بینی می کند. - یک مدل یادگیری ماشینی که برای مورد کاربرد پروژه جاری مناسب است و امکان پیاده سازی دارد.
Feature Engineering	<p>مهندسی ویژگی ها شامل شناسایی نقص، تجمیع و تبدیل متغیرهای خام برای ایجاد ویژگی های مورد استفاده در تجزیه و تحلیل است.</p> <p>اگر می خواهید درک کنید که چه چیزی یک مدل را هدایت می کند، باید بدانید که چگونه ویژگی ها با یکدیگر مرتبط هستند و الگوریتم های یادگیری ماشین برای استفاده از این ویژگی ها چگونه هستند. این مرحله نیاز به ترکیبی خلاق از تخصص در حوزه مربوطه و بینش های به دست آمده از مرحله کاوش داده ها دارد. این مرحله یک عمل متعادل سازی برای یافتن و شامل متغیرهای آموزنده است در حالی که از متغیرهای بیش از حد غیر مرتبط خودداری می کند. متغیرهای مفید و مرتبط نتیجه ما را بهبود می بخشد. متغیرهای غیر مرتبط نویز غیرضروری را به مدل وارد می کنند. همچنین باید برای هر داده جدیدی که در حین امتیازدهی بدست آورده اید، این ویژگی ها را ایجاد کنید. بنابراین تولید این ویژگی ها فقط می تواند به داده هایی بستگی داشته باشد که در زمان اجرای پروژه در دسترس باشد.</p>
Design and Train Model	<p>بسته به نوع سوالی که می خواهید پاسخ دهید، الگوریتم های مدل سازی زیادی در دسترس است. روند طراحی و آموزش مدل شامل مراحل زیر است:</p> <ul style="list-style-type: none"> - داده های ورودی را به طور تصادفی برای مدل سازی به یک مجموعه داده آموزش و یک مجموعه داده آزمون تقسیم کنید. - ساخت مدل ها با استفاده از مجموعه داده های آموزش. - ارزیابی (مجموعه داده های آزمون و آموزش) مجموعه ای از الگوریتم های یادگیری ماشینی رقیب همراه با پارامترهای مختلف تیونینگ مرتبط (معروف به هایپر پارامتر تیونینگ) که جهت پاسخگویی به سوال مورد نظر با داده های فعلی انجام می شود. - "بهترین" راه حل برای پاسخ به سوال را با مقایسه معیار موفقیت بین روش های جایگزین تعیین کنید.

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	 <p>شرکت رایش سریع هزاره ایرانیان www.astek.ir</p>
تهیه کننده: دکتر امین نظارات - شرکت رایش سریع هزاره ایرانیان		www.astek.ir

اجتناب از نشت: نشت داده ها می تواند به دلیل درج داده های خارج از مجموعه داده های آموزشی باشد که به یک الگوریتم یادگیری ماشینی یا مدل اجازه می دهد پیش بینی های غیرواقعی خوبی انجام دهد. نشت دلیل رایجی است که دانشمندان داده را با دریافت نتایج پیشگویی که به نظر می رسد بیش از حد واقعی نیستند، عصبی کند. تشخیص این وابستگی ها دشوار است. برای جلوگیری از این امر معمولاً به تکرار بین ساخت یک مجموعه داده تجزیه و تحلیل، ایجاد یک مدل و ارزیابی دقت نیاز است.

ارزیابی (مجموعه آموزش و مجموعه داده های آزمون و همچنین داده های تست که قبلاً از دیتاست اصلی جدا شده است) مجموعه ای از الگوریتم های یادگیری ماشینی رقیب به همراه پارامترهای مختلف تیونینگ مرتبط که جهت پاسخگویی به سوال مورد نظر با داده های فعلی انجام می شود. اعتبار سنجی فراتر از بررسی کد است. ارزیابی دقیق مفروضات داده ها، پایه کد، عملکرد مدل و نتایج پیش بینی اطمینان را ایجاد می کند که ما می توانیم از طریق علم داده، عملکرد کسب و کار را به طور قابل اعتماد بهبود دهیم. اعتبار سنجی نتایج و تعامل با ذینفعان در این مرحله به همان اندازه مهم است.


برای کشف حالت های خرابی و بهبود تجزیه و تحلیل خطا از خوشه بندی استفاده کنید:

- همه پیش بینی های نادرست را انتخاب کنید.
- یک الگوریتم خوشه بندی مانند دی بی اسکن را در مشاهدات انتخاب شده اجرا کنید
- خوشه ها را به طور دستی جستجو کنید تا به دنبال ویژگی های مشترکی باشید که پیش بینی را دشوار می کند.
- مشاهدات را با پیش بینی های نادرست طبقه بندی کنید و تعیین کنید که برای بهبود عملکرد در این موارد، چه اقدامی می تواند در مرحله ریفاینمنت مدل انجام شود.


Evaluate Model and Discovering failure modes

- از تکرارپذیری و شفافیت پروژه اطمینان حاصل کنید
اعتبار سنجی کیفیت، کالبد شکافی یک مدل و بررسی فرضیه ها و حساسیت ها را از نمونه برداری اولیه تا تنظیم ابر پارامترها و پیاده سازی فرانت اند را در بر می گیرد.


- برای پشتیبانی از بازرسی انسانی از چک های اعتبار سنجی خودکار استفاده کنید
در حالی که ماهیت غیر قطعی علم داده به این معنی است که آزمایش واحد (یونیت تست) به طور مستقیم اعمال نمی شود، اما اغلب مراحل تکرار شده ای در فرآیند اعتبارسنجی وجود دارد که می تواند به صورت خودکار انجام شود. ممکن است مجموعه ای از آمار و نمودارهای خلاصه، آزمایش مجدد نمونه کارها یا هر مرحله دیگری باشد که می تواند به یک

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

	<p>تشخیص خودکار تبدیل شود. این روش اجازه می دهد تا اعتبار سنجی های انسانی بر روی مناطق حساس خاکستری تمرکز کنند.</p> <p>– سوابق داده ها را مورد توجه قرار دهید</p> <p>فرآیند توسعه مدل اغلب به انتخاب ذهنی در مورد پاکسازی داده ها، تولید ویژگی ها و بسیاری از مراحل دیگر نیاز دارد. به عنوان مثال، هنگام ساخت مدل پیش بینی قیمت خدمات، ویژگی "نزدیکی به شعبه" می تواند قدرت پیش بینی را افزایش دهد. با این حال، ممکن است بحث مهم در بین چندین ذینفع در مورد چگونگی محاسبه آن و اینکه آیا از منظر انطباق مجاز است، لازم باشد.</p> <p>– نتایج نال را حفظ کنید</p> <p>حتی اگر پروژه ای هیچگونه ارتقایی به همراه نداشته باشد و به مرحله تولید نرسد مستند ساختن آن و حفظ آن در رپازیتوری بسیار مهم است. غالباً می شنویم که دانشمندان داده ها مشغول انجام کارهایی هستند که کسی بدون اطلاع از سوالات قبلی در آن کاوش کرده است.</p>
Define Feature Sets	<p>ویژگی های توسعه یافته برای مدل سازی در بخش ویژگی های گزارش تعریف داده ها شرح داده شده است. این بخش شامل اشاره گرهای کد برای تولید ویژگی ها و توصیف نحوه تولید ویژگی است.</p>
Add new DataSet to Feature Store	<p>به منظور ایجاد قابلیت استفاده مجدد و همچنین ورژن بندی مجموعه دیتاست های ایجاد شده در پروژه، لازم است که دیتاست به فروشگاه دیتاست ها افزوده شود.</p>
Create Model Report	<p>برای هر مدلی که امتحان می شود، یک گزارش استاندارد و مبتنی بر الگو تهیه می شود که جزئیات هر آزمایش را ارائه می دهد.</p>
Checkpoint Decision	<p>ارزیابی کنید که آیا مدل از عملکرد خوبی برخوردار است تا بتواند آن را در سیستم تولید مستقر کرد. برخی از سوالات اصلی برای پرسیدن عبارتند از:</p> <ul style="list-style-type: none"> – آیا با توجه به داده های آزمون، مدل با اطمینان کافی به این سوال پاسخ می دهد؟ – آیا باید روش های جایگزین دیگری را امتحان کنید: جمع آوری داده های اضافی، مهندسی ویژگی های بیشتر یا آزمایش الگوریتم های دیگر؟

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

Deployment	<p>مدل هایی که دارای پایپ لاین داده هستند برای پذیرش کاربر نهایی در یک محیط تولیدی یا مشابه تولید مستقر می شوند.</p>
Evaluating production readiness	<p>به منظور حصول اطمینان از آمادگی برای عملیاتی کردن نتایج مدل لازم است چک لیست زیر بررسی شود:</p> <p>داده ها:</p> <p>تمامی ویژگیها مورد توجه قرار گرفته اند. همه ویژگی ها مفید هستند. هزینه هیچ ویژگی خیلی زیاد نیست. ویژگی ها به الزامات سطح متا پایبند هستند. خط لوله داده ها دارای کنترل های حریم خصوصی مناسب است. ویژگی های جدید را می توان به سرعت اضافه کرد. همه کدهای دریافت ویژگی ها تست شده اند.</p> <p>مدل:</p> <p>مشخصات مدل بررسی و ارسال می شود. معیارهای آفلاین و آنلاین با هم ارتباط دارند. همه ابر پارامترها تنظیم شده اند. تأثیر ورژنهای گذشته مدل شناخته شده است. یک مدل ساده بهتر نیست؟ کیفیت مدل در برشهای مهم داده کافی است. این مدل برای ملاحظات درج شده آزمایش شده است.</p> <p>زیر ساخت:</p> <p>آموزش قابل تکرار است. مشخصات مورد نیاز مدل از نظر وجودی و کفایت تست شده اند. پایپ لاین یادگیری کامل مدل پیاده سازی شده است. کیفیت مدل قبل از خدمت تأیید می شود. مدل اشکال زدایی شده است. امکان بازگشت به نسخه های مختلف مدل با ای پی های میسر است.</p> <p>مانیتورینگ:</p>

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	
<p>تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان</p>		<p>www.astek.ir</p>

تغییرات اتفاق افتاده در پایپ لاین لاگ شده و هشدار می دهد.
 ناپایدارهای داده ورودی ها را نگهداری می کند.
 ورژن مدل ها قدیمی نیستند.
 مدل ها از نظر عددی پایدار هستند.
 کارایی زیرساخت محاسباتی برای اجرای مدل در طول زمان کافی است.
 دقت پیش بینی از حد آستانه تخطی نکرده است.

هنگامی که مجموعه ای از مدل ها را ارائه می دهید که عملکرد خوبی دارند، می توان آنها را برای استفاده در برنامه های دیگر عملیاتی کرد. بسته به شرایط تجاری، پیش بینی ها به صورت بلادرنگ یا به صورت دسته ای انجام می شوند. برای عملیاتی شدن، مدل ها باید با یک رابط باز در معرض دید مانند ای پی آی (API) که به راحتی از برنامه های مختلف مانند وب سایت های آنلاین، صفحات گسترده، داشبوردها، یا برنامه های تجاری و بک اند قابل دیدن باشند بهره ببرند. همچنین ساختن تله متری و نظارت بر مدل تولید شده و پایپ لاین داده ای که برای کمک به گزارش وضعیت بعدی سیستم و عیب یابی آن استفاده می شود، بهترین روش است.

حتماً یک سیستم نسخه سازی برای موارد زیر داشته باشید:

- پارامترهای مدل
- پیکربندی مدل
- خط لوله ویژه
- مجموعه داده های آموزشی
- مجموعه داده های اعتبار سنجی

از روشهای زیر نیز برای عملیاتی کردن مدل استفاده کنید تا ریسک اجرا کاهش یابد:


:Canarying

ارائه مدل جدید به زیرمجموعه کوچکی از کاربران (یعنی ۵٪) در حالی که هنوز مدل موجود را برای بقیه ارائه نداده اید. این کار به منظور اطمینان از صحت عملکرد مدل و همچنین امکان بازگشت آن به حالت قبلی در صورت بروز خطا می باشد.


:Shadow mode

یک مدل جدید را در کنار مدل موجود سرو کنید، همچنان از مدل موجود برای پیش بینی استفاده کنید اما خروجی هر دو مدل را ذخیره نمایید. اندازه گیری دلتا بین پیش بینی های مدل جدید و فعلی به شما نشان می دهد که چه وقتی به مدل جدید بروید.

Operationalize
the Model

	راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده	
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

	<p>علاوه بر اینها در مرحله عملیاتی کردن مدل لازم است در خصوص موارد زیر نیز تصمیم گیری شود:</p> <ul style="list-style-type: none"> - مشخص کنید چه کسی مسئول خطاهایی است که از پیش بینی های نادرست مدل ناشی می شوند. - جنبه های حقوقی پروژه ها را در مراحل اولیه با یک کارشناس حقوقی در میان بگذارید.
Deliver status dashboard	داشبورد وضعیت سلامتی سیستم و معیارهای کلیدی ایجاد کرده یا توسعه دهید
Deliver final modeling report	گزارش نهایی مدل سازی با جزئیات استقرار را تهیه کنید.
Deliver Final solution architecture document	یک سند راه حل نهایی ایجاد کنید.
Customer Handoff	<p>هدف: نهایی کردن تحویل دانی های پروژه:</p> <p>تأیید کنید که پایپ لاین، مدل و استقرار آنها در یک محیط تولید اهداف مشتری را برآورده می کند.</p>
System Validation	مشتری باید تأیید کند که سیستم نیازهای شغلی آنها را برآورده می کند و با دقت قابل قبولی به سوالات پاسخ می دهد تا سیستم را برای استفاده در محیط عملیاتی وارد کند. تمام اسناد نهایی و بررسی می شود.
Project Handoff	واگذاری پروژه به نهاد مسئول عملیات به پایان رسیده است. این تیم می تواند، به عنوان مثال، یک تیم علم داده سازمان یا نماینده مشتری باشد که مسئولیت اجرای سیستم در تولید را دارد.
Project Close-out: Exit Report of Project for Customer	محصول اصلی تولید شده در این مرحله نهایی، گزارش خروج از پروژه برای مشتری است. این گزارش فنی شامل تمام جزئیات پروژه است که برای یادگیری و کار با سیستم مفید است. یک الگوی گزارش خروج ارائه شده است که می تواند به صورت دلخواه و یا برای نیازهای خاص مشتری سفارشی شود.
Evangelism	یکی از اقدامات مهم در پروژه های علم داده دریافت بازخورد از مشتریان و افزایش میزان دقت مدل بر اساس بازخوردهای دریافتی است. این عملیات کمک می کند که به مرور زمان و پس از دریافت بازخوردهای محیطی، دانشمندان داده بتوانند پارامترهای بهتری برای مدل خود تعیین کنند.
Maintenance Plan	تهیه مستند و اقداماتی که باید در مرحله نگهداری مدل انجام شود در این فعالیت توسط اعضای تیم نهایی می شود و جزئیات این مستند مبنای رابطه کاربران و مجری در فاز نگهداری خواهد بود.

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	 <p>شرکت رایانش سریع هزاره ایرانیان www.astek.ir</p>
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

پروژه های یادگیری ماشینی با ارسال و تحویل نسخه اول کامل نیستند. اگر شما یک پروژه را تحویل می دهید و مسئولیت مدل را منتقل می کنید، بسیار مهم است که برای بهبود و نگهداری مدل برنامه مدونی وجود داشته باشد.

اصول نگهداری

اصل CACE

تغییر هر چیزی همه چیز را تغییر می دهد

۱- سیستم های یادگیری ماشین کاملاً بهم پیوسته اند. تغییر در فضای ویژگی، تغییر در ابر پارامترها، میزان یادگیری یا هر "دستکاری" دیگر می تواند بر عملکرد مدل تأثیر بگذارد.

استراتژی های خاص کاهش اثرات:

آزمون های اعتبار سنجی مدل را ایجاد کنید که با هر بار Push کردن کد جدید اجرا می شود.

در مواردی که منطقی باشد، مشکلات را به اجزای جدا شده تجزیه کنید.

مصرف کنندگان اعلام نشده مدل ممکن است ناخواسته تحت تأثیر تغییرات شما قرار بگیرند.

۲- اگر مدل و/یا پیش بینی های آن به طور گسترده در دسترس باشد، ممکن است سایر اجزای سیستم بدون اطلاع شما به این مدل وابسته شوند. تغییرات در مدل (مانند آموزش مجدد دوره ای یا تعریف مجدد خروجی) ممکن است بر روی اجزای پایین دست تأثیر منفی بگذارد.

استراتژی های خاص کاهش اثرات:

با درخواست مجوز از اجزای خارجی، دسترسی به مدل خود را کنترل کنید تا زمانی که از این مدل استفاده می کنند شما مطلع شوید


۳- از ایجاد وابستگی به سیگنالهای ورودی (داده ها یا پارامترها) که ممکن است با گذشت زمان تغییر کنند، خودداری کنید.

برخی از ویژگی ها با جستجوی جدول (یعنی تعبیه کلمات) یا به سادگی یک پایپ لاین ورودی که خارج از محدوده پایگاه کد شما است، بدست می آیند. با تغییر این ویژگی های خارجی، عملکرد مدل ممکن است آسیب ببیند.


استراتژی های خاص کاهش اثرات:

یک کپی ورژن گذاری شده از سیگنالهای ورودی خود ایجاد کنید تا ثبات را در برابر تغییر در پایپ لاین های ورودی خارجی ایجاد کند. این ورودی های ورژن گذاری شده را می توان در پوشه پیکربندی مدل در دریاچه داده قرار داد.

۴- از بین بردن ویژگی های غیر ضروری

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	
تهیه کننده: دکتر امین نظارات - شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

	<p>به طور منظم اثر حذف ویژگی های فردی از یک مدل داده شده را ارزیابی کنید. فضای ویژگی های یک مدل فقط باید حاوی ویژگی های مهم و برای کار تعیین شده باشد.</p> <p>استراتژی های زیادی برای تعیین اهمیت ویژگی ها وجود دارد، مانند اعتبار سنجی لیو- وان- اوت و آزمون های جایگشت ویژگی. ویژگی های غیر مهم باعث ایجاد نویز به فضای ویژگی شما می شود و باید حذف شود.</p>
<p>Monitoring and Feedback Streaming</p>	<p>فرایند دریافت بازخورد و برچسب زنی بر روی نتایج مدلها در قالب یک خط استریمینگ بین تیم کاربران و مجری می تواند باعث بهبود مستمر مدل شود.</p> <p>نظارت بر کیفیت مدل های علم داده بسیار مهم است. این موضوع به دلیل وابستگی داده ها است. داده ها می توانند با گذشت زمان تغییر کنند، در نتیجه یک مدل علم داده نیز ممکن است با گذشت زمان عملکرد کمتری نشان دهد. به محض تغییر متغیرها و تنظیمات خارجی، داده ها نیز با آن تغییر می کنند. اضافه کردن توضیحی برای پیش بینی یک مدل علم داده، برای کاربر نهایی / کارشناس اجرایی برای درک اینکه این پیش بینی بر اساس چیست، می تواند مفید باشد. این امر بینش و شفافیت بیشتری در مورد چرایی مدل ارائه می دهد</p> <p>تفسیر مدلها:</p> <p>ساختن یک مدل علم داده عمدتاً شامل تکرار است: اجرا و ارزیابی. همچنین در اینجا بازخورد کاربران مهم است. آیا پیش بینی های مدل منطقی است؟ آیا متغیرهای اساسی وجود ندارد که بتواند تأثیرگذار باشد؟ آیا روابط شناسایی شده رابطه علیتی نیز دارند؟ شفافیت الگوریتم نقش مهمی در این روابط دارد. برخی از انواع الگوریتم ها بسیار غیر شفاف هستند، بنابراین نتایج داده شده را نمی توان در داده های ورودی جستجو کرد. این مورد خصوصاً در مورد شبکه های عصبی صادق است. تفسیر روشهای خطی یا رویکردهای آماری سنتی بسیار ساده تر است. تقاضا برای الگوریتم های شفاف را می توان در تحولات اخیر در مورد هوش مصنوعی تفسیرپذیر مشاهده کرد. هنگام انتخاب مدل، با توجه به توضیح پذیری پیش بینی ها، الزامات عملی شفافیت را در نظر بگیرید.</p>
<p>Re- Training Model</p>	<p>از آنجا که میزان داده های دریافتی به طور مستمر در حال افزایش است و مدلهای آموزش دیده باید مجدداً با داده های جدید هم آموزش داده شوند لازم است که این فرایند در کنار اعمال بازخورد در مرحله نگهداری اجرایی شود.</p> <p>وقتی مدلی به طور مکرر بر روی داده های جدید "آموزش مجدد" می شود، مهم این است که به طور مداوم این موارد را ذخیره کنید: چه زمانی، کدام نسخه استفاده می شود، کدام عملکرد متعلق به کدام مجموعه داده آموزشی است. این مورد برای ردیابی عملکرد مدل لازم است.</p>
<p>Accuracy Improving</p>	<p>بررسی میزان بهبود دقت و روالهایی که منجر به این بهبودها شده است می تواند در سایر مدلها و پروژه ها مورد استفاده قرار گیرد.</p>

	<p>راهنمای فعالیتهای مندرج در مستند کنترل پروژه علم داده</p>	
تهیه کننده: دکتر امین نظارات – شرکت رایانش سریع هزاره ایرانیان		www.astek.ir

این امر به دلیل وابستگی شدید بین کدالگوریتم، داده های مورد استفاده برای آموزش الگوریتم و میزان داده های جدید، تحت تأثیر عوامل مختلف خارجی است. ممکن است اتفاق بیفتد که عوامل محیطی تغییر کنند، و در نتیجه فرض خاصی دیگر صحیح نباشد، یا متغیرهای جدیدی ایجاد شوند که قبلاً در دسترس نبودند.

بنابراین توسعه الگوریتم در پس زمینه ادامه می یابد. در نتیجه نسخه های جدیدتر مدل برای همان مورد تجاری با به روزرسانی های نرم افزار ایجاد می شود. در عمل چند نسخه مدل برای مدتی به طور موازی اجرا می شوند، به طوری که تفاوت ها شفاف می شوند. هر مدل وابستگی های خاص خود را دارد که باید مورد توجه قرار گیرد.

در نهایت، پروژه های علوم داده یک حلقه بسته هستند و همیشه پیشرفت ها امکان پذیر است. کد، متغیرها و داده های همیشه در حال تغییر، محصولات علم داده را از یک طرف به لحاظ فنی پیچیده می کند، اما از طرف دیگر بسیار مناسب و قابل استفاده است زیرا به درستی اعمال می شود.

استانداردهایی ایجاد کنید و آنها را برای پروژه های جدید علوم داده مورد استفاده قرار دهید.

- محل ثابت داده (همه در همان محیط).
- ساختار کد سازگار (در صورت توسعه داخلی).
- ساختار داده یکنواخت.