دستورالعمل و قالب حاضر به منظور یکسان­سازی ارائه پایان­نامه­ها تهیه گردیده است. لطفا از این قالب و نرم‌افزار Word استفاده نمایید. به منظور صرفه­جوئی در وقت و اجرای کامل­تر دستورالعمل حاضر، توصیه می­گردد که به جای ساختن یک فایل جدید، با ویرایش فایل متن حاضر پایان­نامه خود را آماده نمایید.

در متن پایان نامه :

* قلم فارسی : B Nazanin، اندازه ۱4، حالت عادی
* قلم لاتین (Latin): Times New Roman، اندازه 12، حالت عادی
* فاصله سطرها 1.5
* تنظیمات پاراگراف: Justified و Right to Left
* برای ایجاد فهرست شکل‌ها، در زبانه REFERENCES به قسمت Insert Table of Figures رجوع شود.
* صفحه عنوان فارسی و انگلیسی تطابق کامل داشته باشد.
* پایان­نامه باید با صفحه‌ی عنوان فارسی شروع شود.
* کلیه روابط در متن شماره به ترتیب شماره گزاری شود.



**پایان‌نامه دوره کارشناسی**

**مهندسی کامپیوتر- گرایش نرم افزار**

عنوان پروژه:

**طراحی و پیاده سازی سیستم تشخیص مدل های یادگیری دانشجویان**

**طراحی و پیاده سازی سیستم پیش بینی نمره دانشجویان**

دانشجو:

پانیذ طاهری

سجاد رحمانی

استاد راهنما:

دكتر کیانیان

**خرداد ۱۴۰۲**



**چکیده**

هوش مصنوعی یکی از علوم مهمی است که امروزه با زندگی ما آمیخته شده است. یکی از حوزه هایی که از این علم استفاده می شود، حوزه درسی و علوم شناختی در مورد مدل یادگیری و پیش بینی نمرات دانشجویان و دانش آموزان است. پیش‌بینی مدل یادگیری و نمرات، یک موضوعی است که تحقیقاتی در مورد آن انجام شده است. محققان به‌دنبال آن هستند که ویژگی هایی که بر یادگیری تاثیر دارند را کشف و با استفاده از آن‌ها مدل یادگیری و نمرات هر دانشجو یا دانش‌آموز را پیش‌بینی کنند. امروز با توجه به پیشرفت علم و فناوری و در دسترس بودن اینترنت برای بیشتر مردم، آموزش مجازی یا LMS خیلی مطرح شده است. محققان سعی می کنند با توجه به فعالیت های هر دانشجو و هر دانش‌آموز در سامانه های یادگیری مجازی مدل یادگیری و نمرات این دانش‌آموز را پیش‌بینی کنند.

**کلمات کلیدی**

مدل یادگیری، نمرات دانشجویان، پیش‌بینی، سامانه یادگیری مجازی، هوش مصنوعی

**فهرست مطالب(این فهرست جهت نمونه آورده شده، فهرست خودتان را با مشورت با استاد راهنما در این قسمت قرار دهید)**

[فصل ۱ 1](#_Toc137326487)

[مقدمه 1](#_Toc137326488)

[۱-۱ مقدمه 1](#_Toc137326489)

[فصل۲ 3](#_Toc137326490)

[پیشینه‌ 3](#_Toc137326491)

[۱-۲ مقدمه 3](#_Toc137326492)

[۲-۲ مدل‌های یادگیری مجازی 3](#_Toc137326493)

[۲-۲-۱ عاطفی 4](#_Toc137326494)

[۲-۲-۲ اجتماعی 4](#_Toc137326495)

[۲-۲-۳ فیزیولوژی 4](#_Toc137326496)

[۲-۲-۴ روانشناسی 5](#_Toc137326497)

[۲-۳ پیش بینی مدل های یادگیری و تحقیقات انجام شده در مورد آن 5](#_Toc137326498)

[۱-۳-۲ The Educational Testing of the Future 6](#_Toc137326499)

[۲-۳-۲ National Assessment of Educational Progress (NAEP) 6](#_Toc137326500)

[۲-۳-۳ Predicting Student Success Using Learning Analytics 7](#_Toc137326501)

[۲-۳-۴ Open University Learning Analytics dataset (OULAD) 7](#_Toc137326502)

[۲-۴ الگوریتم های هوش مصنوعی و پیش‌بینی مدل های یادگیری 8](#_Toc137326503)

[فصل ۳ 10](#_Toc137326504)

[روش انجام کار 10](#_Toc137326505)

[۳-۱ مقدمه 10](#_Toc137326506)

[۳-۲ تحقیق OULAD و Data Set آن 10](#_Toc137326507)

[۳-۲-۱ فایل studetnInfo 10](#_Toc137326508)

[۳-۲-۲ فایل VLE و StudetnVLE 11](#_Toc137326509)

[۳-۲-۳ فایل Assessments و studentAssessment 12](#_Toc137326510)

[۳-۲-۴ فایل Courses و studentRegistration 14](#_Toc137326511)

[۳-۳ فایل data و توضیحات مربوط به آن 15](#_Toc137326512)

[۳-۴ دسته بندی مدل یادگیری دانشجویان 17](#_Toc137326513)

[۳-۴-۱ خوشه بندی داده‌ها به روش K-means 17](#_Toc137326514)

[۳-۴-۲ روش انجام خوشه بندی بر روی داده‌ها 18](#_Toc137326515)

[۳-۵ دسته بندی داده های دانشجویان برای پیش‌بینی نتیجه نهایی 20](#_Toc137326516)

[۳-۵-۱ الگوریتم بیز ساده (Naïve Bayes) 21](#_Toc137326517)

[۲- ۵ -۳ الگوریتم جنگل تصادفی (Random Forest) 22](#_Toc137326518)

[۳-۵-۳ الگوریتم گرادیان افزایشی بسیار قوی (XGBoost) 24](#_Toc137326519)

[۳-۶ مراحل آموزش یک مدل (model training) 25](#_Toc137326520)

[۱-۶-۳ جمع آوری داده 25](#_Toc137326521)

[۲-۶-۳ پیش پردازش داده‌ها 25](#_Toc137326522)

[۳-۶-۳ تقسیم داده‌ها به دو مجموعه آموزشی و آزمایشی 27](#_Toc137326523)

[۴-۶-۳ طراحی مدل 28](#_Toc137326524)

[۵-۶-۳ آموزش مدل 29](#_Toc137326525)

[۳-۷ ارزیابی عملکرد مدل(evaluating the performance of machine learning models) 31](#_Toc137326526)

[۱-۷-۳ معیار ارزیابی دقت 32](#_Toc137326527)

[۲-۷-۳ معیار دقت صفر 34](#_Toc137326528)

[۳-۷-۳ ماتریس اغتشاش 35](#_Toc137326529)

[۴-۷-۳ احتمالات چند کلاسه 36](#_Toc137326530)

[فصل 4 نتیجه‌گیری 40](#_Toc137326531)

[4-1 مقدمه 40](#_Toc137326532)

[مراجع 41](#_Toc137326533)

[واژه نامه انگلیسی به فارسی 43](#_Toc137326534)

[پیوست‌ها 44](#_Toc137326535)

فهرست جدول‌ها

[جدول ‏2‑1: نتایج حاصل از سنتز دو مدار 46](#_Toc111099205)

فهرست شکل‌ها

شکل ‏2‑1: ورودی و خروجی های DHT22 **Error! Bookmark not defined.**

# فصل ۱

# مقدمه

## ۱-۱ مقدمه

پژوهش در مورد بهبود یادگیری دانش‌آموزان و دانشجویان از موضوعاتی بوده که در مورد آن تحقیقات زیادی انجام شده و نتایج خوبی هم کسب شده. برای این امر محققان و پژوهشگران درصدد تقسیم بندی مدل های یادگیری هستند و سپس تولید محتوا برای دانشجو و دانش‌آموز با توجه به مدل یادگیری اوست. این تحقیقات سعی کرده با کسب اطلاعاتی همچون ارزیابی ها و فعالیت دانشجویان و دانش‌آموزان مدل های یادگیری را پیش بینی کند؛ همچنین با استفاده از این اطلاعات سعی کردند که نمره دانشجویان را پیش بینی کنند. این تحقیق ها با پرسشنامه های ساده شروع شده و امروز با استفاده از روش های یادگیری ماشین و داده کاوی و اطلاعات سامانه های یادگیری مجازی (LMS) ، مدل یادگیری هر دانشجو و نمره او را پیش‌بینی می‌کنند.

یکی از این تحقیقات در Open University (دانشگاهی در بریتانیا) با بیش از ۳۲ هزار دانشجو انجام شد. این پژوهش سعی کرده بود با توجه به فعالیت ها و ارزیابی هایی که دانشجویان در سامانه یادگیری مجازی انجام داده اند ، و روش های یادگیری ماشین و داده کاوی مدل یادگیری و موفقیت دانشجویان را پیش بینی کند.

پیش‌بینی مدل یادگیری و میزان موفقیت دانشجویان به بهبود کیفیت آموزش، مدیریت بهتر کلاس‌های آنلاین، بهبود نرخ موفقیت دانشجویان، بهره‌وری بیشتر از منابع آموزشی و بهبود تجربه دانشجویان کمک می‌کند. با پیش‌بینی مدل یادگیری دانشجو می‌توانیم محتوایی را در اختیار او قرار دهیم که به یادگیری بهتر وی ختم شود. برای مثال اگر کسی مدل یادگیری‌اش دیداری هست، برای یادگیری بهتر وی می‌توان محتوای تصویری تهیه کرد. با این کار یادگیری دانشجو یا دانش‌آموز بهتر می‌شود و می تواند نتایج بهتری کسب کند.

در این پژوهش ما سعی کردیم با استفاده از دیتاست پژوهش OULAD و با استفاده از الگوریتم های مختلف یادگیری ماشین پیش بینی میزان موفقیت دانشجویان و دسته بندی مدل یادگیری دانشجویان را بدست آوریم. در این پژوهش قصد داشتیم میزان دقت هر الگوریتم را بسنجیم تا بهترین الگوریتم را برای این کار انتخاب کنیم. علاوه بر این برای پیش بینی مدل یادگیری دانشجویان باتوجه به میزان فعالیت هر دانشجو در سامانه مجازی آنها را با استفاده از روش k-means به ۵ گروه تقسیم کردیم.

# فصل۲

# پیشینه‌

## ۱-۲ مقدمه

مدل‌های یادگیری مختلفی برای یادگیری آموزشی در مدارس و دانشگاه ها وجود دارد و هر فردی مدل یادگیری مخصوص به خود را دارد. همین علت باعث شده که محقیقن و پژوهش‌گران با انجام پژوهش‌های مختلف در صدد پیش بینی مدل یادگیری دانشجویان و پیش‌بینی موفقیت دانشجویان هستند. این تحقیقات از سال ۱۹۶۰ با پرسش‌نامه‌های دستی شروع شده و و با ثبت فعالیت‌های هر دانشجو در سامانه‌های LMS ادامه دارد.

## ۲-۲ مدل‌های یادگیری مجازی

یادگیری سنتی و یادگیری آنلاین از بسیاری جهات متفاوت است. به عنوان مثال، دانش آموزان در یک محیط کلاس درس ممکن است ترجیحاتی در مورد صدا، نور و دما داشته باشند، در حالی که این عوامل برای محیط های یادگیری آنلاین مناسب نیستند زیرا عنصر اصلی محیط یک صفحه وب تعاملی است. بنابراین، یک مدل سبک یادگیری آنلاین از هشت بعد در کار قبلی ما پیشنهاد شده است. این اثر یادگیری سنتی و آنلاین را در چهار دسته عاطفه، جامعه شناسی، فیزیولوژی و روانشناسی مقایسه می کند. پس از آن، هشت ویژگی برای مشخص کردن یادگیرندگان آنلاین طراحی شده است، و بررسی رفتارهای یادگیری آنلاین مرتبط با این ویژگی ها انجام و تجزیه و تحلیل می شود. نتایج نشان می‌دهد که مدل جدید سبک یادگیری آنلاین ما، زبان‌آموزان آنلاین را متمایز می‌کند و به درک رفتار آنها کمک می‌کند.

### ۲-۲-۱ عاطفی

مقوله عواطف حول محور این است که یادگیرندگان آنلاین تا چه اندازه یادگیرندگان خودراهبر هستند. بر اساس مدل Entwistle، یادگیرندگان آنلاین با انگیزه خود را تا پایان دوره نظارت می‌کنند و سرعت می‌گیرند، بنابراین ممکن است سوابق تعاملی بیشتری با سیستم‌های آموزش الکترونیکی داشته باشند و تمایل دارند که بر منابع یادگیری نامحبوب نسبت به همتایان بی‌انگیزه خود کلیک کنند. در مقابل، یادگیرندگان غیرفعال به سادگی مطالب و ارزیابی های لازم را به پایان می رسانند

### ۲-۲-۲ اجتماعی

یادگیرندگان آنلاین همچنین در نحوه واکنش آنها به تعامل و ارتباط با همسالان متفاوت هستند. برخی بحث را دوست ندارند و ترجیح می دهند خودشان مطالعه کنند. دیگران در حمایت از کار گروهی رشد می کنند

### ۲-۲-۳ فیزیولوژی

ویژگی های بصری و کلامی به مدل فلدر-سیلورمن اشاره دارد. زبان آموزان آنلاین اطلاعات را از منابع مختلف دریافت می کنند: دیداری (مانند مناظر، تصاویر، نمودارها و نمادها) و شنیداری (مانند صداها و کلمات). یادگیرندگان دیداری از نظر بصری حساس تر هستند و درک بهتری از مطالب ارائه شده به صورت دیداری دارند، در حالی که زبان آموزان شنوایی با گوش دادن یا خواندن مطالب، اطلاعات را با عملکرد بهتر به دست می آورند. سایر فراگیران با نحوه ارائه مطالب سازگار می شوند.

### ۲-۲-۴ روانشناسی

مقوله روانشناسی به راهبردهایی اشاره دارد که دانش آموزان برای درک اطلاعات از آنها استفاده می کنند. ویژگی‌های حسی و شهودی که به نشانگر نوع مایرز-بریگز (MBTI) اشاره می‌کند، منعکس‌کننده چیزی است که فراگیران توجه خود را روی آن متمرکز می‌کنند. یادگیرندگان حسی مطالب دقیق را بر اساس حقایق ترجیح می دهند، در حالی که یادگیرندگان شهودی مفاهیم، معانی و تداعی ها را ترجیح می دهند. علاوه بر این، ویژگی‌های متوالی و کلی را از مدل فلدر-سیلورمن معرفی می‌کنیم زیرا ترتیب ارائه مواد بر کارایی یادگیری تأثیر می‌گذارد. برخی از آنها به طور متوالی در یک پیشرفت منطقی منظم یاد می گیرند، و برخی دیگر با جهش های شهودی یاد می گیرند تا در نهایت بفهمند. در اکثر مدل های سنتی، ویژگی ها متقابلاً منحصر به فرد هستند. به عنوان مثال، در مدل فلدر-سیلورمن، یک یادگیرنده نمی تواند همزمان متوالی و سراسری باشد. برعکس، مدل ما آن ویژگی ها را با استفاده از بردار هشت بعدی برای مشخص کردن زبان آموزان ترکیب می کند. به عنوان مثال، اگر یادگیرنده نمرات بالایی در هر دو ویژگی متوالی و کلی کسب کند، می توان گفت که یادگیرنده در انتخاب راهبردهای یادگیری بسیار انعطاف پذیر است.

## ۲-۳ پیش بینی مدل های یادگیری و تحقیقات انجام شده در مورد آن

تحقیقات زیادی برای پیش‌بینی مدل یادگیری انجام شده است. از سال ۱۹۶۰ اولین تحقیق رسمی شروع شد و تاکنون این تحقیقات ادامه دارد. برای انجام این پژوهش‌ها از هوش مصنوعی هم کمک گرفته شده است که با استفاده از این علم بتوان مدل‌های یادگیری دانشجویان و دانش‌آموزان را بهتر و دقیق تر پیش‌بینی کرد. در ادامه چند نمونه از تحقیقات انجام شده را بیان کردیم. هدف از بیان این تحقیقات سیر تغییرات و پیشرفت و افزایش دقت در انجام این پژوهش ها بود.

### ۱-۳-۲ The Educational Testing of the Future

این تحقیق در سال ۱۹۶۰ توسط Russell Ackoff و Fred Emery در مدارس بریتانیا انجام شد. این تحقیق به دنبال آن بود که با برگزاری یک آزمون شامل ۶۰ سوال، دانش آموزان را به ۴ دسته تقسیم کند:

۱. دانش آموزانی که به خوبی در موضوعات مختلف عملکرد خوبی داشتند.

۲. دانش آموزانی که در موضوعات خاصی عملکرد بالایی داشتند، اما در موضوعات دیگر بهترین عملکرد را نداشتند.

۳. دانش آموزانی که در موضوعات مختلف عملکرد متوسطی داشتند.

۴. دانش آموزانی که در بیشتر موضوعات عملکرد ضعیفی داشتند.

این دسته‌بندی، به دانش آموزان و معلمان کمک می‌کرد تا نقاط قوت و ضعف دانش آموزان را شناسایی کنند و برای هر دانش آموز، برنامه‌ی آموزشی مناسبی را طراحی کنند. این تحقیق، به عنوان یک پایه برای تحقیقات بعدی در زمینه مدل‌سازی و پیش‌بینی موفقیت دانش آموزان شناخته شده است.

### ۲-۳-۲ National Assessment of Educational Progress (NAEP)

National Assessment of Educational Progress یا NAEP یک برنامه ملی آزمون است که به منظور ارزیابی سطح آموزش و پیشرفت دانش آموزان در ایالات متحده آمریکا طراحی شده است. این برنامه تحت نظارت سازمان National Center for Education Statistics (NCES) قرار دارد.

NAEP برای ارزیابی دانش آموزان در سطوح مختلف آموزشی، از طریق اجرای آزمون‌های استانداردی استفاده می‌کند. این آزمون‌ها شامل سوالات چندگزینه‌ای و پاسخ کوتاه هستند و به منظور ارزیابی مهارت‌های خواندن، نوشتن، ریاضی و علوم اجرا می‌شوند.

### ۲-۳-۳ Predicting Student Success Using Learning Analytics

پژوهش "Predicting Student Success Using Learning Analytics" یکی از مطرح‌ترین تحقیقات در حوزه یادگیری و تحلیل داده‌های آموزشی است. در این پژوهش، با استفاده از تحلیل داده‌هایی که از لرنینگ مدیریت سیستم‌های (LMS) بدست آمده بود، سعی شده است تا با پیش‌بینی موفقیت یا شکست دانشجویان، راهکارهایی برای بهبود عملکرد و پیشرفت دانشجویان ارائه شود.

در این پژوهش، از الگوریتم‌های یادگیری ماشینی و تحلیل داده‌های آموزشی برای پیش‌بینی موفقیت یا شکست دانشجویان استفاده شده است. با استفاده از این الگوریتم‌ها، تحلیل‌هایی بر روی داده‌هایی از دانشجویان انجام شده و با پیش‌بینی موفقیت یا شکست آن‌ها، راهکارهایی برای بهبود عملکرد و پیشرفت دانشجویان ارائه شده است.

هدف این پژوهش ارائه راهکارهایی برای بهبود کارایی و پیشرفت دانشجویان بوده است. به عنوان مثال، با تحلیل داده‌های آموزشی، می‌توان مشخص کرد که فرآیند یادگیری دانشجویان در کدام مرحله مشکل دارد و با ارائه راهکارهایی متناسب با آن مشکل، بهبود عملکرد و پیشرفت دانشجویان را تسریع کرد.

از دیگر کاربردهای این پژوهش می‌توان به بهبود فرآیند تدریس و طراحی درس‌ها، بهبود روش‌های ارزیابی دانشجویان و همچنین ارائه بازخورد دقیق‌تر به دانشجویان اشاره کرد.

## ۲-۳-۴ Open University Learning Analytics dataset (OULAD)

تحقیقات متعددی در دانشگاه باز Open University UK و به ویژه در پروژه‌ی OULAD (Open University Learning Analytics Dataset) انجام شده است که به پیش‌بینی موفقیت و شکست دانشجویان در دوره‌های آموزشی این دانشگاه پرداخته‌اند.

این پروژه، داده‌های جمع‌آوری‌شده از دانشجویانی که در دوره‌های آموزشی Open University UK شرکت کرده‌اند را شامل می‌شود و از آنها برای پیش‌بینی موفقیت دانشجویان استفاده شده است. در این پروژه از روش‌های یادگیری ماشین، شبکه‌های عصبی و الگوریتم‌های داده‌کاوی برای پیش‌بینی موفقیت دانشجویان استفاده شده است.

در این پروژه، از داده‌های مختلفی مانند اطلاعات شخصی دانشجویان، تعداد بار ورود به سامانه، تعداد بازدید از دروس، نمرات در آزمون‌ها و اطلاعات زمانی استفاده شده است. برای هر دانشجو، یک پروفایل آموزشی تهیه شده و با استفاده از این پروفایل، میزان موفقیت دانشجو در دوره‌ی آموزشی شان پیش‌بینی شده است.

این تحقیقات نشان داده‌اند که با استفاده از داده‌های جمع‌آوری‌شده از دانشجویان، می‌توان به صورت دقیق‌تری پیش‌بینی موفقیت دانشجویان در دوره‌های آموزشی شان را انجام داد و در نتیجه، بهبود کیفیت آموزش و یادگیری را به دنبال داشت.

## ۲-۴ الگوریتم های هوش مصنوعی و پیش‌بینی مدل های یادگیری

امروزه، مطرح‌ترین علمی که با زندگی ما آمیخته شده است «هوش مصنوعی» است. از صنعت گرفته تا پزشکی، همگی به نوعی از هوش مصنوعی استفاده می‌کنند. از این علم در پیش‌بینی مدل‌های یادگیری هم استفاده می‌شود. الگوریتم های زیادی از هوش مصنوعی در این امر استفاده می‌شود که می‌خواهیم به ۴ مورد از پراستفاده ترین این الگوریتم‌ها اشاره کنیم:

* شبکه‌های عصبی (Neural Networks): این الگوریتم به طور گسترده‌ای در مسائل پیش‌بینی مورد استفاده قرار می‌گیرد. این شبکه‌ها معمولاً شامل لایه‌های مختلفی از نورون‌ها هستند که با هم متصل شده‌اند و قادر به یادگیری الگوهای پیچیده هستند.
* درخت تصمیم (Decision Trees): این الگوریتم به صورت درختی عمل می‌کند و برای پیش‌بینی بر اساس شرایط مختلف استفاده می‌شود. درخت تصمیم شامل گره‌ها و شاخه‌هایی است که با توجه به شرایط مختلف، به یکی از دو شاخه مختلف تعلق می‌گیرند.
* کلاسی فایرهای بیزی (Bayesian Classifiers): این الگوریتم برای پیش‌بینی و تشخیص الگوهای پیچیده با استفاده از روش‌های احتمالاتی استفاده می‌شود. در این الگوریتم، احتمال پیشین و احتمال شرطی برای پیش‌بینی و مدل‌سازی استفاده می‌شود.
* الگوریتم ایکس جی بوست (XGBoost) : XGBoost یکی از محبوب‌ترین الگوریتم‌های یادگیری ماشینی برای پیش‌بینی و مدل‌سازی است. XGBoost مخفف "eXtreme Gradient Boosting" است و در واقع یک الگوریتم ترکیبی از چندین درخت تصمیم (decision tree) است که با استفاده از روش Gradient Boosting کار می‌کند.

ما در این پژوهش از الگوریتم random forest (یکی از الگوریتم های درخت تصمیم)، naïve bayes و XGboost استفاده کردیم. در هر قسمت معیار‌های ارزیابی را بررسی کردیم که بهترین الگوریتم را برای این کار پیدا کنیم.

**نتیجه‌گیری**

پیش‌بینی مدل یادگیری یکی از عواملی است که با شناخت بهتر و دقیق تر آن می‌توان به دانش‌آموزان و دانشجویان برای یادگیری بهتر، کمک کرد. پژوهش هایی برای این امر از سال ۱۹۶۰ تا کنون انجام شده است و همگی این پژوهش ها سعی می‌کنند دقیق تر مدل یادگیری را مشخص کنند. هر چه تحقیق ها جدید تر شدند، ویژگی‌های بیشتری از فعالیت های دانشجویان (مخصوصا در سامانه های یادگیری مجازی) و نمراتی که از ارزیابی ها گرفتند، بررسی می‌شوند که هم مدل یادگیری مرتبط با دانشجو به اون پیشنهاد دهیم و هم میزان موفقیت این دانشجو را مشخص کند. الگوریتم های هوش مصنوعی مختلفی برای انجام این کار استفاده می‌شود که ما در پی یافتن دقیق ترین و بهینه ترین آن‌ها برای این کار هستیم.

# فصل ۳

# روش انجام کار

## ۳-۱ مقدمه

در این بخش قصد داریم با بررسی فعالیت‌های هر دانشجو در تحقیق OULAD ، مدل یادگیری دانشجو را دسته بندی و همچنین میزان موفقیت وی را پیش‌بینی کنیم. برای این امر ابتدا اطلاعات این تحقیق را شرح داده و سپس بالگوریتم‌هایی کخ برای پیش‌بینی میزان موفقیت دانشجو انجام دادیم را معرفی می‌کنیم و در نهایت هر کدام را ارزیابی کردیم.

## ۳-۲ تحقیق OULAD و Data Set آن

تحقیق OULAD برای پیش‌بینی میزان موفقیت دانشجو و تحلیل و تجزیه رفتار دانشجویان انجام شده است. این تحقیق بر روی ۳۲۵۹۳ نفر در ۲۲ دوره انجام شد. ۱۰۶۵۵۲۸۰ فعالیت در سامانه VLE از دانشجویان ثبت گردیده. این فعالیت‌ها را به انواع مختلفی تقسیم کرده اند که دانشجو با کلیک بر روی هر رویدادی، یکی از فعالیت‌ها برای وی در نظر گرفته می‌شود. اطلاعات جمع آوری شده در این تحقیق در ۸ فایل وجود دارد. در ادامه هر جدول را به طور خلاصه با داده هایی که دارد معرفی می‌کنیم.

### ۳-۲-۱ فایل studetnInfo

این فایل شامل اطلاعات دانشجویان در دوره‌های آموزشی مختلف است، به شکل یک جدول با ستون‌های مختلف. برخی از ستون‌های موجود در این فایل عبارتند از:

جدول ۱-۳ . اطلاعات موجود در جدول studentInfo

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Id\_student | شناسه دانشجو |
| Gender | جنسیت دانشجو |
| Region | منطقه محل سکونت دانشجو |
| Highest\_education | سطح تحصیلات |
| Age\_band | بازه سنی دانشجو |
| Num\_of\_prev\_attempts | تعداد بار های شرکت در یک دوره |
| Studied\_credits | تعداد واحد هایی که دانشجو در این درس گذرانده |
| Disability | آیا دارای معلولیت است یا خیر؟ |

### ۳-۲-۲ فایل VLE و StudetnVLE

فایل VLE شامل اطلاعات سامانه یادگیری مجازی و تقسیم بندی فعالیت‌های موجود در آن است. برای اتصال این فایل و دانشجویان به یکدیگر ما از فایل studentVLE استفاده می‌کنیم که این دو فایل به وسیله id\_site و id\_student به یکدیگر وصل شده اند. یعنی در فایل studentVLE متوجه می‌شویم هر دانشجو در چه روز چه تعدادی یک فعالیت را در سامانه یادگیری مجازی انجام داده است. فایل studentVLE و VLE شامل اطلاعات زیر هستند:

جدول ۲-۳. اطلاعات موجود در جدول VLE

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Id\_site | شناسه سایت که منحصر به فرد است |
| Code\_module | کد دوره |
| Code\_presention | کد تاریخ برگزاری |
| Activity\_type | نوع فعالیت در سامانه |
| Week\_from | هفته شروع برای فعالیت |
| Week\_to | تا هفته چندم این فعالیت بوده |

جدول ۳-۳. اطلاعات موجود در جدول studentVLE

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Id\_site | شناسه سایت که منحصر به فرد است |
| Code\_module | کد دوره |
| Code\_presention | کد تاریخ برگزاری |
| Date | زمان انجام فعالیت |
| Sum\_click | تعداد انجام فعالیت |

### ۳-۲-۳ فایل Assessments و studentAssessment

فایل assessments شامل تمام ارزیابی ها انجام شده است. این ارزیابی ها در سه قالب انجام شده است:

* ارزیابی هایی که توسط استاد انجام شده است (TMA)
* ارزیابی هایی که توسط سیستم انجام شده است (CMA)
* امتحان ها (EXAM)

بقیه اطلاعات این فایل در جدول ۴-۳ اشاره شده است. برای اتصال فایل ارزیابی ها و دانشجویان بهم ما از فایل studentAssessment استفاده کردیم که در این فایل با استفاده از id\_assessment و id\_student می‌توانیم نمره هر دانشجو در هر ارزیابی را بدست بیاوریم. اطلاعات این فایل هم در جدول ۵-۳ وجود دارد.

جدول ۴-۳. اطلاعات موجود در جدول Assessments

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Id\_assessment | شناسه ارزیابی |
| Code\_module | کد دوره |
| Code\_presention | کد تاریخ برگزاری |
| Assessment\_type | نوع ارزیابی که سه حالت دارد |
| Date | تعداد روز هایی که از شروع دوره گذشته |
| Weight | میزان تاثیر ارزیابی در نمره پایانی |

جدول ۵-۳. اطلاعات موجود در جدول studentAssessment

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Id\_assessment | شناسه ارزیابی |
| Id\_student | شناسه دانشجو |
| Is\_banked | آیا نمره ذخیره شده یا خیر |
| Date\_submitted | تعداد روز هایی که طول کشیده تا تایید شود |
| Score | نمره نهایی |

### ۳-۲-۴ فایل Courses و studentRegistration

در فایل Courses ما تمام دوره هایی که انجام شده را داریم. همان طور که ذکر شد، این تحقیق در ۲۲ دوره انجام شده است. هر دوره یک code\_presentation دارد که ابتدا سال را بیان می‌کند و کاراکتر بعدی ماه را مشخص می‌کند. برای مثال 2013J به معنی این هست که دوره از اکتبر (ماه دهم) سال ۲۰۱۳ شروع شده است.اطلاعات این فایل در جدول ۶-۳ قرار داده شده است. همچنین برای ارتباط بین این فایل و فایل دانشجویان فایلی به نام studentRegistration ایجاد شده است که اطلاعات این فایل در جدول ۷-۳ ذکر شده است.

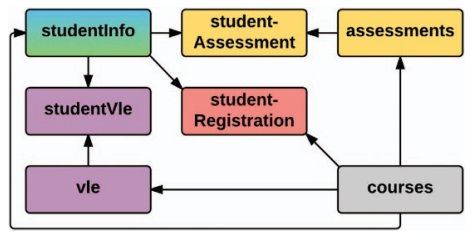
جدول ۶-۳. اطلاعات موجود در جدول Courses

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Code\_module | کد دوره |
| Code\_presention | کد تاریخ برگزاری |
| module\_presentation\_length | مدت دوره به تعداد روز |

جدول ۷-۳. اطلاعات موجود در جدول studentRegistration

|  |  |
| --- | --- |
| **عنوان** | **توضیح** |
| Code\_module | کد دوره |
| Code\_presention | کد تاریخ برگزاری |
| Id\_student | شناسه دانشجو |
| Date\_registration | تاریخ ثبت نام دانشجو |
| Date\_unregistration | تاریخ لغو ثبت نام دانشجو |

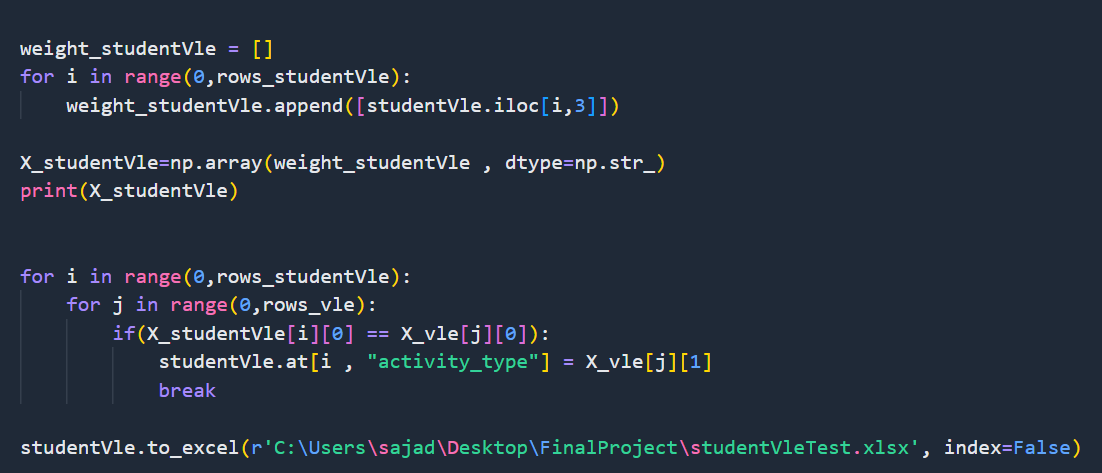
می‌توان تمام فایل ها و اطلاعات این پژوهش را به صورت شکل ۱-۳ نشان داد.



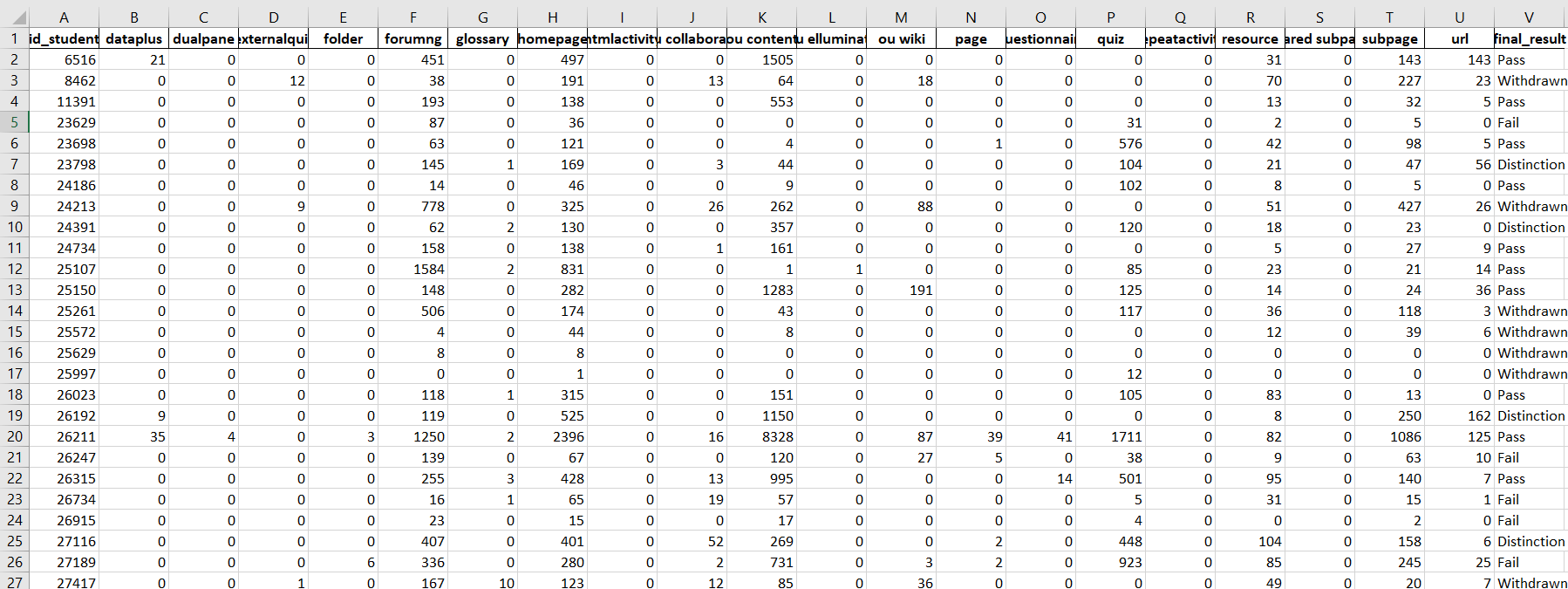
شکل ۱-۳. ارتباط بین اطلاعات و فایل های تحقیق OULAD

### ۳-۳ فایل data و توضیحات مربوط به آن

ما در این پژوهش سعی کردیم با استفاده از فعالیت‌های دانشجویان در سامانه VLE مدل یادگیری آن‌ها را تقسیم بندی کنیم و سپس با توجه به اطلاعات آن میزان موفقیت یک دانشجو را پیش‌بینی کنیم. برای این امر ما با استفاده از فایل VLE تمام نوع فعالیت ها را در فایل StudentVle قرار می‌دهیم. این کار را با کمک کد شکل ۲-۳ انجام شد. بعد از این کار ما با استفاده از id\_student که در فایل studentVle وجود دارد برای هر نوع فعالیت، تمام کلیک های دانشجو را با هم جمع کردیم و در فایل دیتا ما اطلاعات را به صورت شکل ۳-۳ داریم.



شکل ۲-۳ . کد اضافه کردن نوع فعالیت در فایل StudentVleTest.



شکل ۳-۳ . نمایی از فایل data که با استفاده از این فایل قرار است مدل یادگیری را دسته بندی و میزان موفقیت دانشجو را پیش بینی کنیم.

برای اینکه هیچ خانه ای خالی نماند، اگر دانشجویی در یک نوع فعالیت خاص هیچ رکوردی ثبت نکرده باشد در داخل آن خانه عدد صفر قرار می‌دهیم. همچنین فایل دیتا شامل ۲۶۰۷۵ داده هست در حالی که ۳۲۵۹۳ دانشجو داریم. علت این تفاوت به دو دلیل زیر است:

* در فایل studentInfo ما تعداد id\_student تکراری داشتیم که حدود ۳۸۰۰ تا بودند.
* حدود ۲۸۰۰ id\_student ، هیچ تراکنش و فعالیت ثبت شده ای داخل فایل studentVle نداشتند.

## ۳-۴ دسته بندی مدل یادگیری دانشجویان

یکی از اهداف ما در این پژوهش این بود که با توجه به فعالیت های دانشجویان در سامانه VLE مدل یادگیری آن‌ها را خوشه‌ی کنیم. ما از روش K-means برای خوشه بندی استفاده کردیم. در این قسمت قصد داریم ابتدا روش K-means را توضیح دهیم و سپس روش اعمال این روش بر روی داده‌های خودمان را شرح دهیم.

### ۳-۴-۱ خوشه بندی داده‌ها به روش K-means

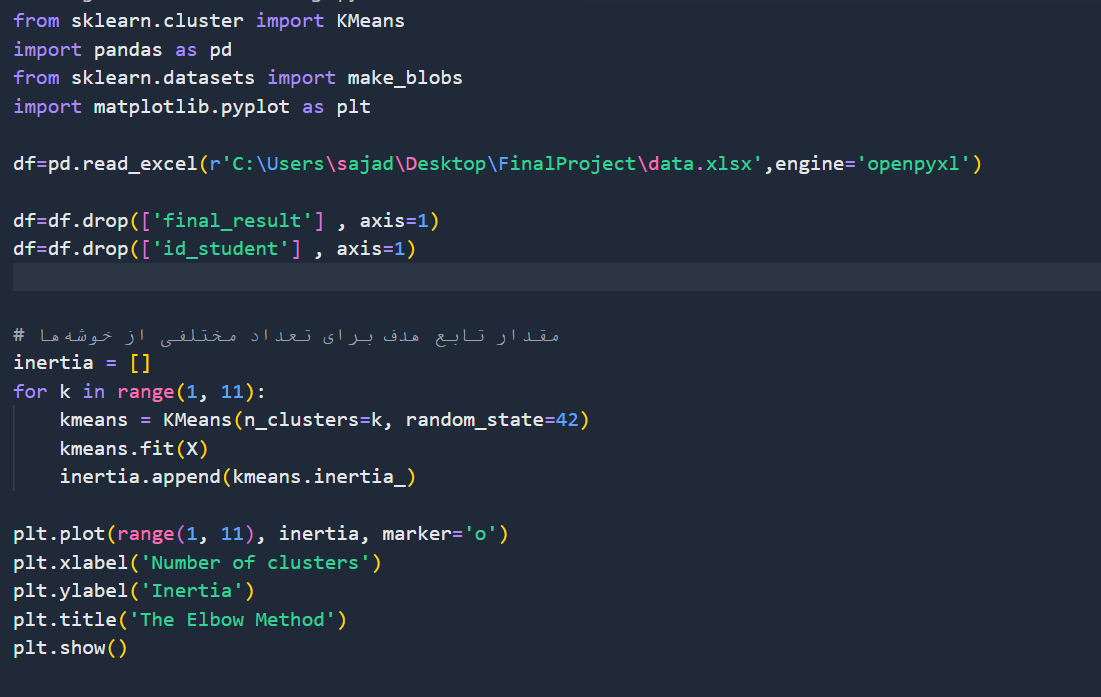
خوشه‌بندی k-means یک روش کوانتیزه‌سازی برداری است که در اصل از پردازش سیگنال است و هدف آن تقسیم n مشاهده به k خوشه است که در آن هر مشاهده متعلق به خوشه‌ای با نزدیک‌ترین میانگین (مراکز خوشه یا مرکز خوشه) است که به عنوان نمونه اولیه از خوشه این منجر به پارتیشن بندی فضای داده به سلول های Voronoi می شود[].در خوشه‌بندی K-means ما داده ها را براساس فاصله اقلیدسی از مرکز خوشه، خوشه‌بندی می‌کنیم. یعنی برای اینکه بدانیم یک داده را در یکی از خوشه‌ها قرار دهیم، فاصله آن داده را تا مرکز هر خوشه حساب می‌کنیم و سپس خوشه‌ای را انتخاب می‌کنیم که داده ما تا مرکز آن خوشه کمترین فاصله را داشته باشد. برای انجام این الگوریتم مراحل زیر را انجام می‌دهیم[]:

* تعداد خوشه های K را مشخص کنید.
* ابتدا با به هم زدن مجموعه داده ها و سپس انتخاب تصادفی K نقاط داده برای مرکز خوشه‌ها بدون جایگزینی، مرکز خوشه‌ها را راه اندازی کنید.
* به تکرار ادامه دهید تا زمانی که تغییری در مرکز خوشه‌ها ایجاد نشود. یعنی تخصیص نقاط داده به خوشه ها تغییر نمی کند.

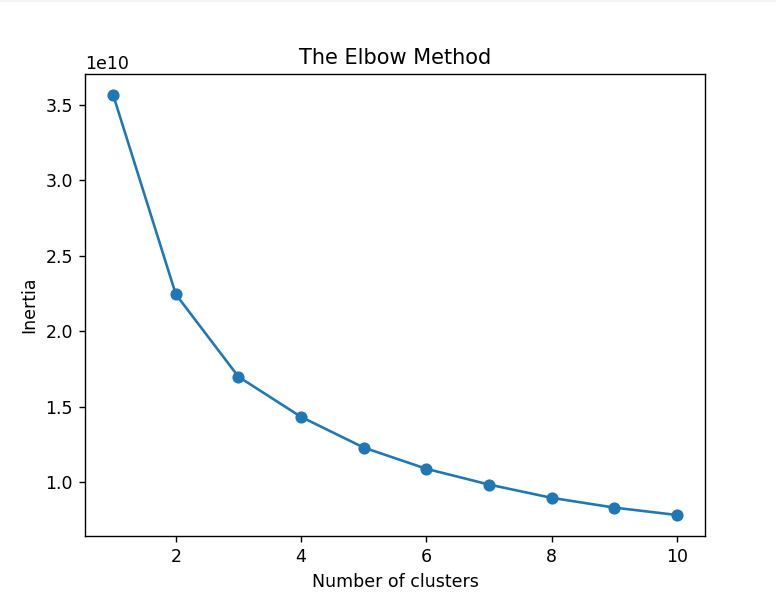
برای محاسبه فاصله هر داده از مرکز خوشه، از فرمول اقلیدسی استفاده می‌کنیم:

### ۳-۴-۲ روش انجام خوشه بندی بر روی داده‌ها

همان‌طور که ذکر شده بود ما تعداد تراکنش‌های یک دانشجو را در دوره‌های مختلف بدست آوردیم و در فایل data قرار دادیم. با توجه به این اطلاعات می‌خواهیم مدل یادگیری دانشجویان را خوشه‌بندی‌ کنیم. اما یک سوال را در ابتدا باید برای خودمان پاسخ دهیم: اینکه تعداد خوشه‌های ما چندتاست؟ برای تعیین تعداد خوشه‌ها ما از روش L-bow استفاده کردیم. برای این کار ما با استفاده از کد شکل ۴-۳، نمودار را رسم کردیم و با توجه به نمودار(شکل ۵-۳) بهترین مقدار برای تعداد خوشه‌ها عدد ۵ بود.

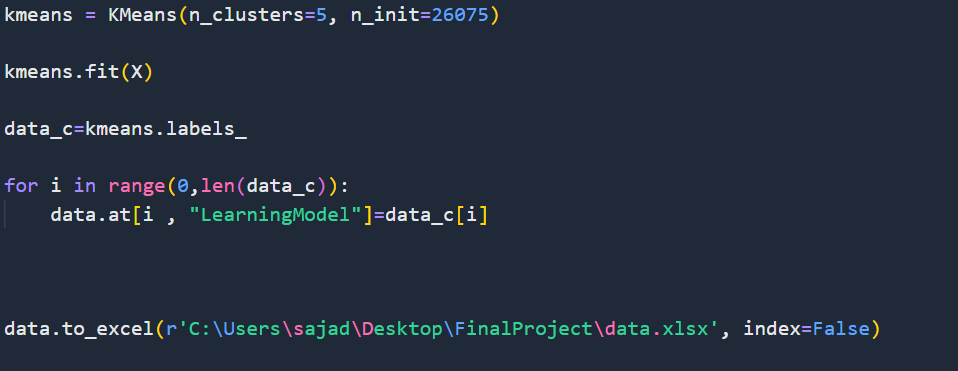


شکل ۴-۳. مقدار خطا ها را برای تعداد خوشه‌های مختلف با استفاده از این کد حساب شده.

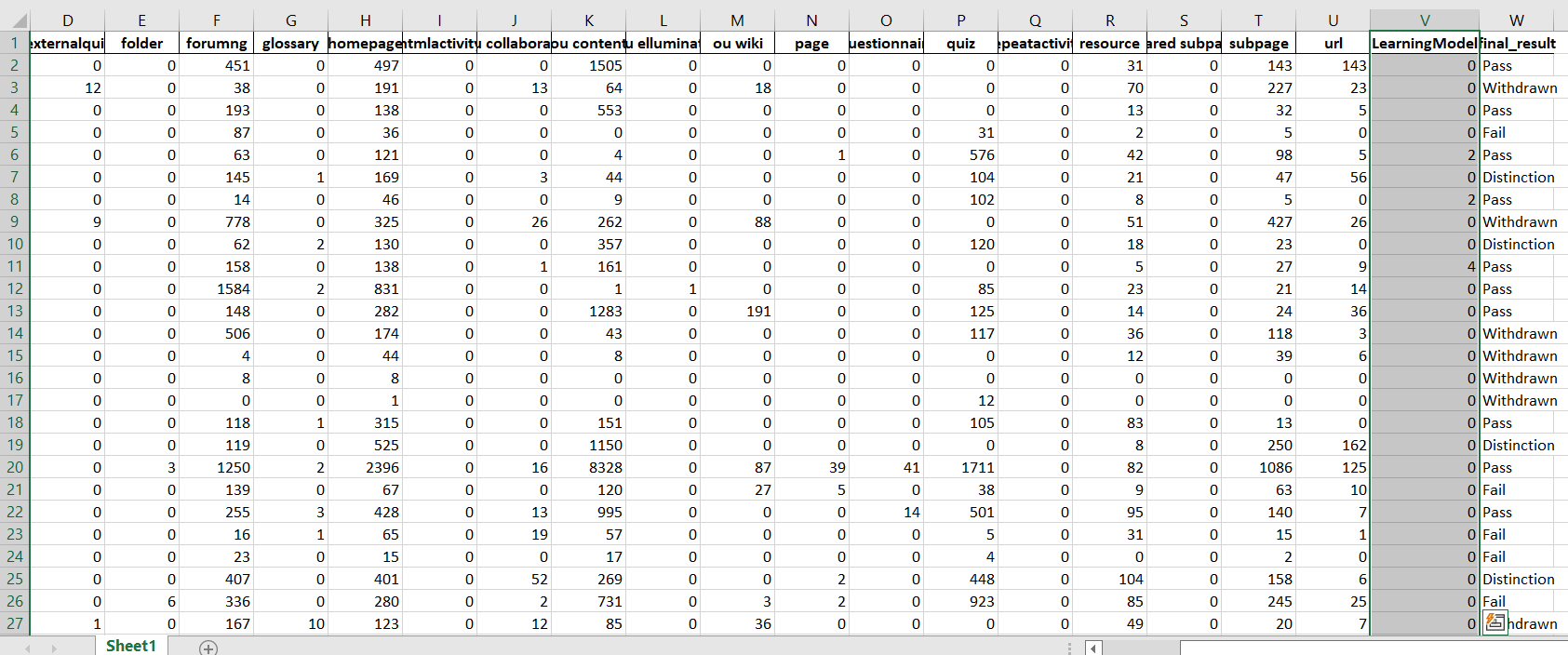


شکل ۵-۳ . نمودار L-bow که باتوجه به نمودار شکستگی در نقطه k=5 است.

با توجه به اینکه برای ما مشخص شد که تعداد خوشه‌ها چندتاست پس با توجه به این مورد حال کد مربوط به خوشه‌بندی را به صورت شکل ۶-۳ می‌زنیم.



شکل ۶-۳. کد خوشه بندی داده ها با تعداد خوشه ۵.



شکل ۷-۳. نتیجه اجرای کد شکل ۶-۳ که در فایل data مدل‌های یادگیری خوشه بندی شدند و برچسب 0 الی 4 به داده ها تعلق گرفته است.

## ۳-۵ دسته بندی داده های دانشجویان برای پیش‌بینی نتیجه نهایی

از مهمترین مفاهیم مورد کاربرد ما در این پژوهش مفهوم طبقه‌بندی(classification) در یادگیری ماشین(machine learning) است. در واقع طبقه بندی یکی از روش‌های یادگیری ماشین است که در آن، یک سیستم یادگیری ماشین به دنبال دسته‌ای از نمونه‌ها می‌گردد و سپس هر نمونه را به یکی از دسته‌های موجود در داده‌های آموزشی نسبت می‌دهد. در اینجا نتیجه نهایی(Final\_Result) دانشجویان ما دارای ۴ دسته می باشد (Fail , Pass , Withdrawn , Distinction)

به‌ طور کلی، در یک مسئله طبقه‌بندی، هدف این است که سیستم یادگیری ماشین بتواند دسته های موجود را بهتر تمیز دهد. در اینجا هدف تشخیص نتیجه نهایی برای هر داشجو است. برای این کار، در ابتدا داده‌های آموزشی به سیستم یادگیری ماشین ارائه می‌شود. سپس سیستم یادگیری ماشین بر اساس ویژگی‌هایی که از داده‌های آموزشی استخراج می‌کند، یک مدل آموزش داده می‌شود. این مدل به عنوان یک تابعی عمل می‌کند که برای هر نمونه جدیدی که به سیستم وارد می‌شود، مقداری را­ به عنوان خروجی تولید می‌کند که نشان دهنده دسته‌ای است که نمونه جدید به آن تعلق دارد. که در همین مقاله به نحوه ی آموزش مدل آن با برخی از الگوریتم ها و استفاده از در زبان پایتون خواهیم پرداخت.

برای طبقه بندی الگوریتم های متنوعی وجود دارد :

* الگوریتم درخت تصمیم‌گیری(Decision Tree)
* الگوریتم بیز ساده (Naïve Bayes)
* الگوریتم جنگل تصادفی (Random Forest)
* الگوریتم ماشین بردار پشتیبان (SVM)
* الگوریتم شبکه‌های عصبی
* الگوریتم‌های رده‌بندی بایاس-واریانس (Bias-Variance)
* الگوریتم گرادیان افزایشی بسیار قوی(XGBoost)

در اینجا می‌خواهیم به سه الگوریتمی که در این پژوهش استفاده کردیم بیشتر بپردازیم و مورد بررسی قرار دهیم:

* الگوریتم بیز ساده (Naïve Bayes)
* الگوریتم جنگل تصادفی (Random Forest)
* الگوریتم گرادیان افزایشی بسیار قوی(XGBoost)

### ۳-۵-۱ الگوریتم بیز ساده (Naïve Bayes)

این الگوریتم بر اساس قانون بیز که یک اصل مهم در آمار و احتمالات محاسباتی است، عمل می‌کند. اساس این قانون این است که با دانستن وقوع یک رویداد، احتمال وقوع یک رویداد دیگر محاسبه می‌شود. در الگوریتم بیز ساده، برای دسته‌بندی داده‌ها، از احتمالات شرطی که بر اساس داده‌های ورودی محاسبه می‌شوند، استفاده می‌شود. به‌طور ساده، الگوریتم بیز ساده داده را به دسته‌ای که احتمال بیشتری دارد، تخصیص می‌دهد. برای محاسبه احتمالات شرطی در الگوریتم بیز ساده، از فرضیات ساده‌ای استفاده می‌شود که به‌عنوان فرض ساده شناخته می‌شود. این فرضیه این است که ویژگی‌های ورودی به‌طور مستقل از هم هستند و احتمال وقوع هر ویژگی به‌صورت جداگانه محاسبه می‌شود و سپس احتمال شرطی دسته‌ها برای هر ویژگی با استفاده از قانون بیز محاسبه می‌شود.

قانون بیز:

که در آنP(A|B) نشان‌دهنده احتمال وقوع حادثه A با توجه به رخ دادن B است.  
P(B|A) نشان‌دهنده احتمال وقوع B در صورت رخ دادن حادثه A است.  
P(A) نشان‌دهنده فرضیه اولیه در مورد احتمال وقوع حادثه A است.  
P(B) نشان‌دهنده احتمال وقوع B است.

### ۲- ۵ -۳ الگوریتم جنگل تصادفی (Random Forest)

الگوریتم جنگل تصادفی یا Random Forest Algorithm یک الگوریتم یادگیری ماشینی است که برای مسائل کلاس بندی و رگرسیون استفاده می‌شود. این الگوریتم از ترکیب چند درخت تصمیم‌گیری (Decision Tree) تشکیل شده است و به دلیل قابلیت یادگیری در داده‌های بزرگ و پایداری در برابر انواع داده‌های نویزی به عنوان یکی از الگوریتم‌های پرکاربرد در حوزه یادگیری ماشینی شناخته می‌شود.در الگوریتم جنگل تصادفی، ابتدا تعدادی درخت تصمیم‌گیری با استفاده از روش دسته‌بندی CART (Classification and Regression Trees) ساخته می‌شود. به‌طور کلی، هر درخت تصمیم‌گیری برای پیش‌بینی خروجی، مجموعه‌ای از قوانین در قالب یک درخت گرافیکی را ارائه می‌دهد. سپس، برای هر نمونه ورودی، تمامی درخت‌های تصمیم‌گیری روی داده‌های آموزشی آموزش داده شده، اعمال می‌شوند و خروجی هر درخت برای هر نمونه محاسبه می‌شود. در نهایت، خروجی نهایی برای هر نمونه به دست می‌آید که از طریق تصمیم‌گیری روی خروجی‌های درخت‌ها به‌دست می‌آید.

با استفاده از روش جنگل تصادفی، مشکل برازش بیش‌ازحد (Overfitting) که در الگوریتم درخت تصمیم‌گیری وجود دارد، کاهش می‌یابد. این روش با تولید تعداد زیادی از درخت‌های تصمیم‌گیری با استفاده از داده‌های آموزشی، از انجام برازش بیش‌ازحد خودداری می‌کند و همچنین، با ترکیب خروجی‌های مختلف درخت‌های تصمیم‌گیری، دقت پیش‌بینی را افزایش می‌دهد. با توجه به اینکه الگوریتم جنگل تصادفی یک الگوریتم یادگیری ماشینی پیشرفته است، در ادامه به بیان بیشتر جزئیات مراحل آن می پردازیم:

1. ساخت درخت‌های تصمیم‌گیری: در ابتدا، چندین درخت تصمیم‌گیری ساخته می‌شود که هر کدام از آن‌ها به صورت مستقل از داده‌های آموزشی ساخته می‌شوند.

۲. انتخاب نمونه‌های تصادفی: برای ساخت هر درخت تصمیم‌گیری، برخی از نمونه‌های داده‌های آموزشی به صورت تصادفی انتخاب می‌شوند. این کار باعث می‌شود که درخت‌های تصمیم‌گیری مستقل از هم باشند.

3. ساخت درخت‌های تصمیم‌گیری با روش دسته‌بندی CART: هر درخت تصمیم‌گیری با استفاده از روش دسته‌بندی CART ساخته می‌شود. در این روش، برای ساخت هر درخت، به صورت تصادفی یک ویژگی و یک مقدار آن انتخاب می‌شود و سپس برای تقسیم داده‌های آموزشی، به دو دسته براساس آن ویژگی تقسیم می‌شوند. این پروسه تا رسیدن به برگ‌های درخت تصمیم‌گیری ادامه می‌یابد. هر برگ درخت تصمیم‌گیری یک تصمیم نهایی برای پیش‌بینی خروجی ارائه می‌دهد.

4. پیش‌بینی خروجی با استفاده از درخت‌های تصمیم‌گیری: برای پیش‌بینی خروجی برای هر نمونه ورودی، تمامی درخت‌های تصمیم‌گیری روی داده‌های آموزشی آموزش داده شده، اعمال می‌شوند و خروجی هر درخت برای هر نمونه محاسبه می‌شود. در نهایت، خروجی نهایی برای هر نمونه به دست می‌آید که از طریق تصمیم‌گیری روی خروجی‌های درخت‌ها به‌دست می‌آید.

### ۳-۵-۳ الگوریتم گرادیان افزایشی بسیار قوی (XGBoost)

 XGBoost (eXtreme Gradient Boosting) یک الگوریتم پر قدرت و کارآمد برای حل مسائل طبقه‌بندی و رگرسیون است که بر پایه الگوریتم Gradient Boosting ایجاد شده است. XGBoost به طور خاص برای پردازش مجموعه داده‌های بزرگ با ویژگی‌های با ابعاد بالا طراحی شده است. الگوریتم Gradient Boosting با اضافه کردن یادگیرنده‌های ضعیف (معمولاً درخت‌های تصمیم) به مدل کار می‌کند، به طوری که هر یادگیرنده جدید با جبران خطاهای باقیمانده از یادگیرنده قبلی سازگار می‌شود. در XGBoost، یک عبارت انحرافی اضافه می‌شود به تابع هدف برای جلوگیری از بیش‌برازش. علاوه بر این،XGBoost از تکنیک نمونه‌برداری مبتنی بر گرادیان برای انتخاب زیرمجموعه‌ای از ویژگی‌های اطلاعاتی برای هر درخت استفاده می‌کند، که هزینه محاسباتی را کاهش می‌دهد و دقت مدل را افزایش می‌دهد. XGBoost دارای چندین مزیت نسبت به سایر الگوریتم‌های یادگیری ماشین است، از جمله:

1. دقت بالا: XGBoost نشان داده است که در بسیاری از مجموعه داده‌های بنچ‌مارک، عملکرد برتری دارد.
2. قابلیت مقیاس‌پذیری: XGBoost می‌تواند با مجموعه داده‌های بزرگ با میلیون‌ها نمونه و هزاران ویژگی کار کند.
3. کارآیی: XGBoost برای بهینه‌سازی بسیار بهینه شده است و مدل‌ها را با سرعتی بسیار بالاتر از دیگر پیاده‌سازی‌های Gradient Boosting آموزش می‌دهد.
4. پایداری: XGBoost قادر به کار با مقادیر گمراه‌کننده، پرت‌ و سایر نواحی داده‌ای نامنظم است.
5. انعطاف‌پذیری: XGBoost می‌تواند برای حل مسائل رگرسیون و طبقه‌بندی استفاده شود و می‌تواند با انواع مختلفی از داده‌ها از جمله متن، تصاویر و ویژگی‌های دسته‌ای کار کند.

XGBoost در حوزه‌های مختلفی مانند مالی، بهداشت، و پردازش زبان طبیعی به گسترش رسیده است. برخی از کاربردهای رایج XGBoost شامل تشخیص تقلب، مدل‌سازی ریسک اعتباری، تشخیص سرطان، و تحلیل احساسات هستند. به طور کلی XGBoost یک الگوریتم یادگیری ماشین قدرتمند و انعطاف‌پذیر است که می‌تواند به راه‌حل‌های دقیق و کارآمد برای یک طیف گسترده‌ای از مسائل پیش‌بینی‌ای منجر شود.

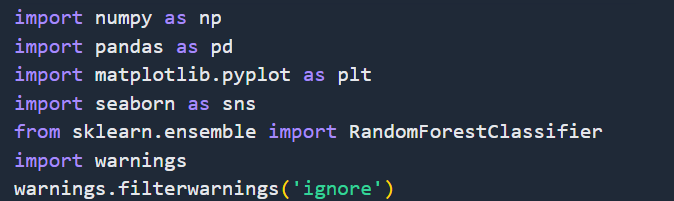
## ۳-۶ مراحل آموزش یک مدل (model training)

### ۱-۶-۳ جمع آوری داده

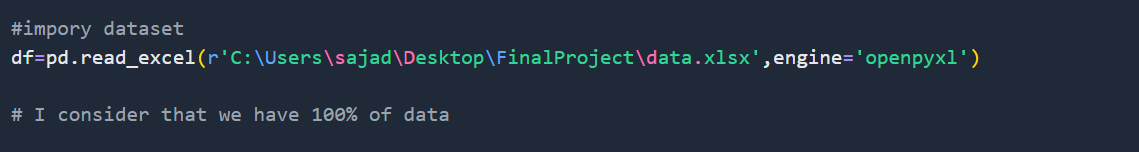
اولین مرحله در آموزش یک مدل ماشین، جمع‌آوری داده‌های مربوط به مسئله است که قرار است مدل به آنها آموزش داده شود. این داده‌ها می‌توانند از منابع مختلفی مانند پایگاه داده‌ها، فایل‌های متنی، تصاویر و ویدئوها جمع‌آوری شوند. که در این پژوهش همانطور که قبلا توضیح داده شد از مجموعه داده بنچ‌مارک OULAD برای استخراج داده ها استفاده شده است.

### ۲-۶-۳ پیش پردازش داده‌ها

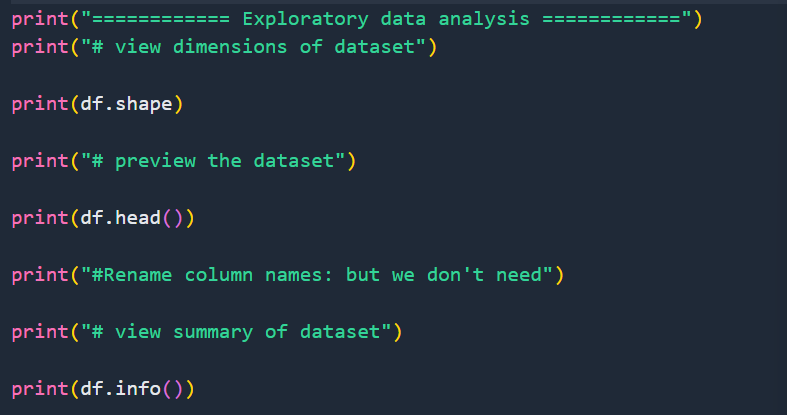
در این مرحله، داده‌های جمع‌آوری شده برای آموزش مدل به شکلی مناسب تبدیل می‌شوند. این شامل انجام کارهایی مانند پاکسازی داده‌ها، حذف داده‌های ناقص و تبدیل داده‌های متنی به بردارهای عددی است که مدل بتواند با آنها کار کند. در تکه کد های زیر بعد از وارد(import) کردن کتابخانه ها و همینطور داده ی مورد نظر پیش پردازش(analysis) روی آن صورت می گیرد.



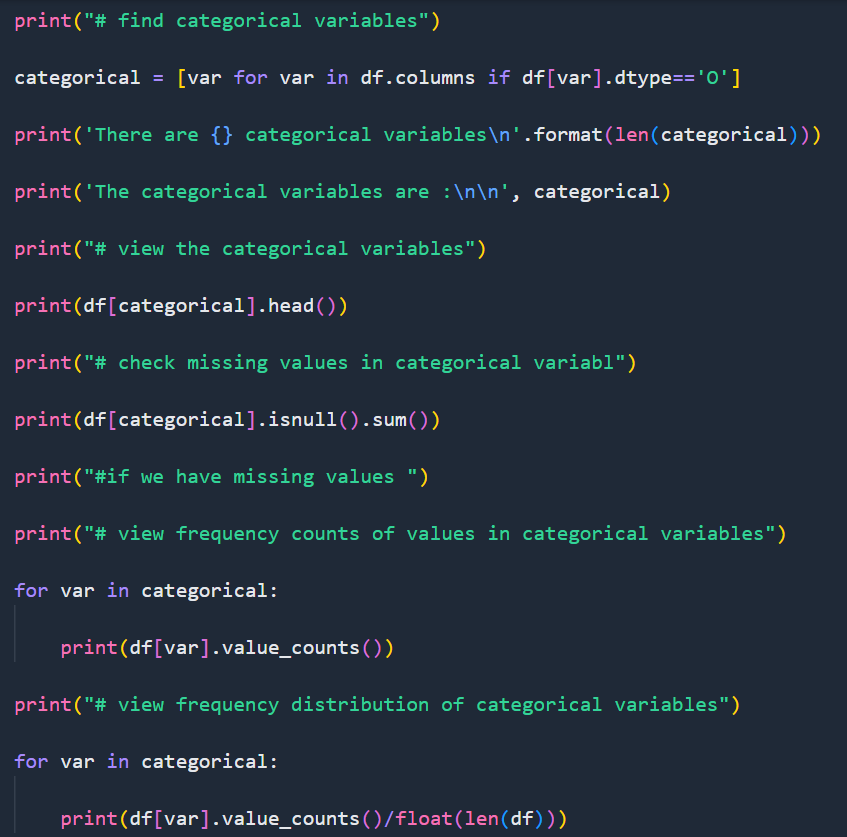
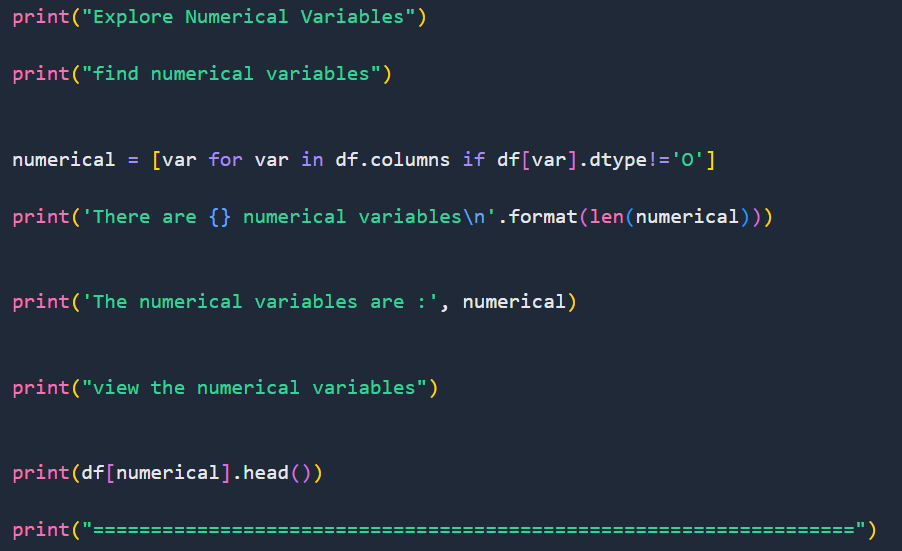
شکل ۸-۳. کد افزودن کتابخانه های لازم در این پروژه در زبان پایتون



شکل ۹-۳. کد افزودن فایل اکسل لازم در این پروژه در زبان پایتون.



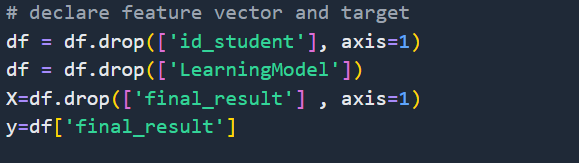
شکل ۱۰-۳. نمایش اطلاعات کلی جدول داده ها در زبان پایتون

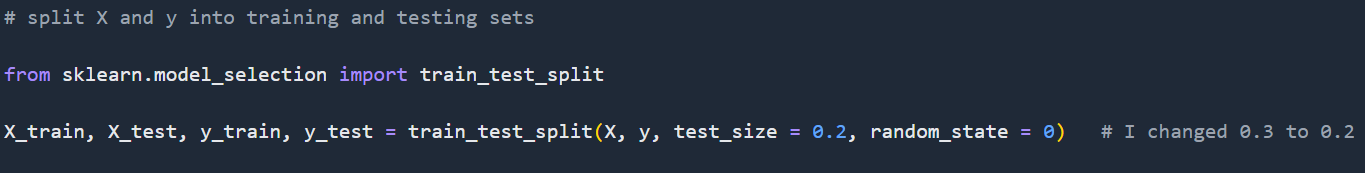
شکل ۱۱-۳ کد بررسی ویژگی‌های غیر عددی در زبان پایتون

### ۳-۶-۳ تقسیم داده‌ها به دو مجموعه آموزشی و آزمایشی

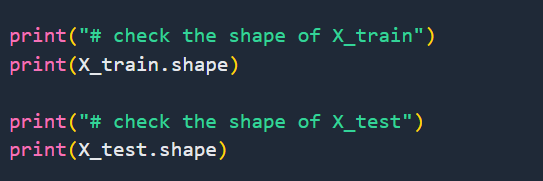
برای ارزیابی عملکرد مدل، داده‌ها به دو مجموعه تقسیم می‌شوند. مجموعه آموزشی برای آموزش مدل استفاده می‌شود و مجموعه آزمایشی برای ارزیابی عملکرد مدل استفاده می‌شود. مراحل این فرایند به ترتیب در تکه کد های زیر نمایش داده شده است.



شکل ۱۲-۳. جدا کردن داده های ورودی و حذف برچسب‌های اضافی از فایل data در پایتون

شکل ۱۳-۳ جدا کردن داده های ورودی و برچسب های آموزشی و آزمونی به ۸۰ به ۲۰ درصد در پایتون

train\_test\_split در scikit-learn یک تابع بسیار مفید برای تقسیم داده‌ها به دو بخش آموزش(train) و آزمون(test) است. این تابع از بسته model\_selection در scikit-learn برای تقسیم داده‌های یک مجموعه داده به دو بخش استفاده می‌شود. در شکل ۳-۱۳ تابع train\_test\_split با گرفتن داده‌های ورودی x و برچسب‌ها y و درصد تقسیم آن‌ها بین داده‌های آموزش و آزمون، داده‌ها را به صورت تصادفی به دو بخش تقسیم می‌کند.



شکل ۱۴-۳. بررسی ابعاد داده های ورودی آموزشی و آزمونی در پایتون

### ۴-۶-۳ طراحی مدل

در این مرحله، مدل ماشین برای حل مسئله طراحی می‌شود. این شامل انتخاب نوع مدل (مانند بیز ساده. شبکه‌های عصبی، درخت تصمیم و... ) و تعیین پارامترهای مدل (مانند تعداد لایه‌ها، تعداد نورون‌ها و سایر پارامترهای مشخص کننده مدل) است. در این قسمت ما تصمیم گرفتیم از مدل بیز ساده (naïve bayes) نام برده استفاده کنیم.

### ۵-۶-۳ آموزش مدل

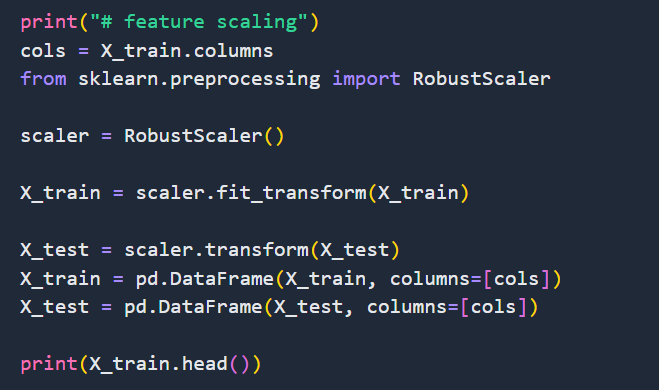
در این مرحله، مدل با استفاده از داده‌های آموزشی آموزش داده می‌شود. این شامل اعمال الگوریتم یادگیری به داده‌های آموزشی و بهبود عملکرد مدل با انجام چرخه‌های آموزش است. نکته ای بسیار مهم که قبل از آموزش مدل باید در نظر داشت مقیاس بندی ویژگی ها(Feature scaling) می باشد. مقیاس بندی ویژگی ها در پیش‌بینی مهم است، این شامل بسیاری از الگوریتم‌های یادگیری ماشین است که برای پیش‌بینی استفاده می‌شوند. به طور کلی، مقیاس بندی ویژگی ها به فرایند نرمال‌سازی یا استاندارد سازی مقادیر ویژگی‌های ورودی اشاره دارد تا آن‌ها در یک مقیاس مشابه قرار بگیرند. این می‌تواند باعث بهبود عملکرد مدل یادگیری ماشین شود با اطمینان از اینکه ویژگی‌ها به طور مساوی مهم هستند و یکدیگر را به خود نمی‌کشند. تکنیک‌های مختلفی برای مقیاس بندی ویژگی ها وجود دارند که شامل استانداردسازی، مقیاس‌بندی min-max و RobustScaler,Robust scaling می‌شوند. به طور کلی، پیشنهاد می‌شود که قبل از آموزش مدل یادگیری ماشین، به‌ویژه در الگوریتم‌هایی که حساس به مقیاس ویژگی‌های ورودی هستند، مانند K-Nearest Neighbors، Support Vector Machines و Neural Networks، مقیاس بندی ویژگی‌ها انجام شود. در اینجا ما از روش RobustScaler برای استاندارد سازی داده ها استفاده کرده ایم. RobustScalerیک روش استانداردسازی است که در برابر داده‌های پرت و نویزدار مقاوم است.

معمولا در استانداردسازی داده‌ها، میانگین و واریانس داده‌ها در نظر گرفته می‌شوند و داده‌ها با توجه به این میانگین و واریانس، به یک مقیاس استاندارد نگاشت می‌شوند. با این حال، این روش در برابر داده‌های پرت و نویزدار ضعیف عمل می‌کند و ممکن است خروجی پرتی در برگرداند.

در RobustScaler، به جای استفاده از میانگین و واریانس، از مقادیر میانه و رنج بین کوچکی داده‌ها استفاده می‌شود. به طور دقیق‌تر، در این روش، ابتدا برای هر ویژگی (ستون) از داده‌ها، مقدار میانه و رنج بین کوچکی داده‌ها (با کمک مقدار ۲۵ و ۷۵ درصدی) محاسبه می‌شود. سپس، با استفاده از فرمول زیر، داده‌ها به مقیاس استاندارد نگاشت می‌شوند:

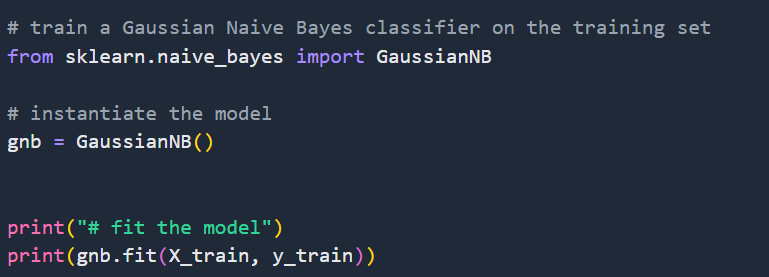
که در آن، x به داده مورد نظر اشاره دارد، median میانه داده‌های ویژگی مربوطه، و IQR برابر با رنج بین کوچکی داده‌های ویژگی است.

استفاده از RobustScaler به دلیل مقاومت بیشتر این روش نسبت به روش‌های سنتی استانداردسازی، به ویژه در صورت داشتن داده‌های پرت و نویزدار، توصیه می‌شود.

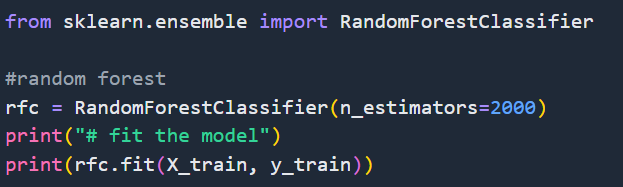


شکل ۱۵-۳. مقیاس بندی ویژگی ها با RobustScaler در پایتون

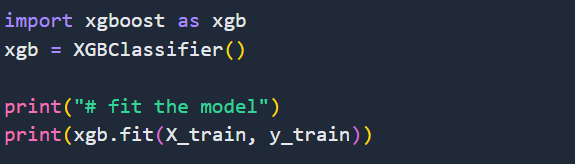
تا اینجا ما داده ها را آماده کردیم. حال می‌خواهیم که سه الگوریتمی که گفته شد را در این قسمت قرار دهیم. کد این سه الگوریتم در شکل های ۱۶-۳، ۱۷-۳ و ۱۸-۳ زده شده است.



شکل۱۶-۳. مدل آموزشی naïve bayes در پایتون

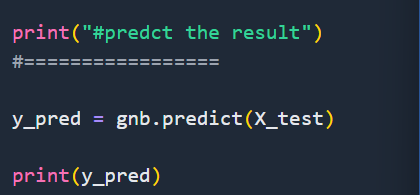


شکل۱۷-۳. مدل آموزشی random forest در پایتون



شکل۱۸-۳. مدل آموزشی XGBoost در پایتون

حال مدل ما آماده ی پیش بینی final\_result دانشجویان(fail,pass, distinction, withdrawn) است. در تکه کد زیر برای داده های آزمون (X\_test) پیش بینی صورت گرفته است.



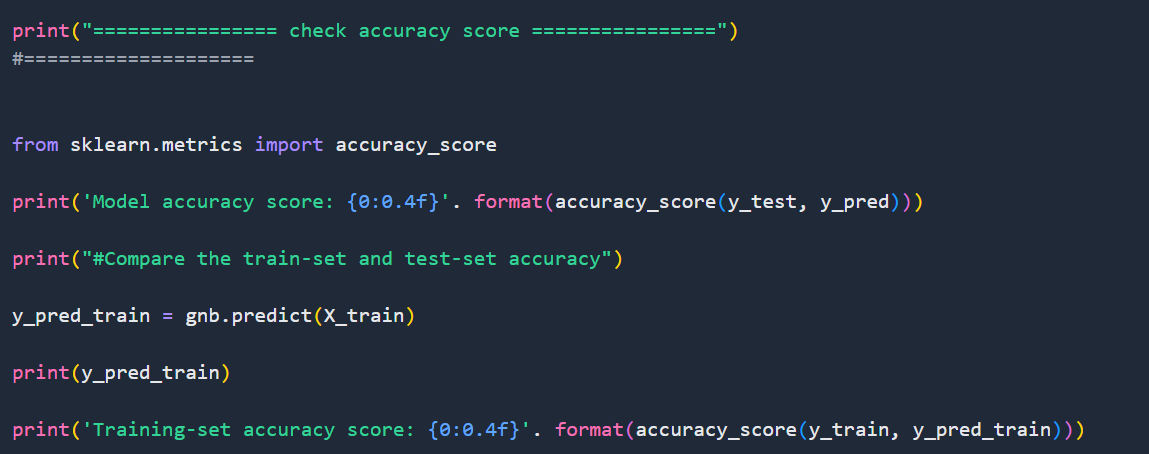
شکل ۱۹-۳. پیش بینی برچسب های داده های آزمونی در پایتون

## ۳-۷ ارزیابی عملکرد مدل(evaluating the performance of machine learning models)

پس از اتمام آموزش مدل، عملکرد آن با استفاده از داده‌های آزمایشی ارزیابی می‌شود. این شامل محاسبه معیارهای ارزیابی مانند دقت، صحت و سایر معیارهای مشخص کننده عملکرد مدل است. معیارهای مورد استفاده برای ارزیابی عملکرد مدل ممکن است بسته به نوع مسئله و نوع داده‌ها متفاوت باشد. برای مثال، در مسئله این پژوهش که پیش‌بینی دسته‌بندی است، معیارهایی مانند دقت (accuracy)، صحت (precision)، بازخوانی (recall) و امتیاز F1-score استفاده می‌شود. معمولاً ماتریس درهم‌ریختگی (confusion matrix) نیز برای ارزیابی دقیق‌تر عملکرد مدل استفاده می‌شود. که در ادامه توضیحات بیشتری داده خواهد شد.

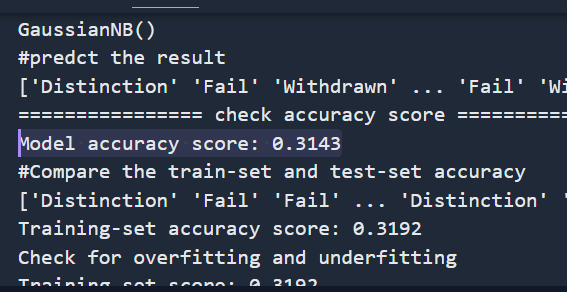
### ۱-۷-۳ معیار ارزیابی دقت (accuracy)

این معیار نشان دهنده تعداد نمونه‌هایی است که به درستی توسط مدل دسته‌بندی شده‌اند. به عبارت دیگر، دقت برابر است با تعداد نمونه‌هایی که به درستی تشخیص داده شده‌اند تقسیم بر تعداد کل نمونه‌ها.با استفاده از تابع accuracy\_score از ماژول sklearn.metrics می‌توان دقت را محاسبه کرد. تکه کد پایتون زیر دقت(accuracy) پیش بینی را برای داده های آزمون و آموزش نمایش می دهد.



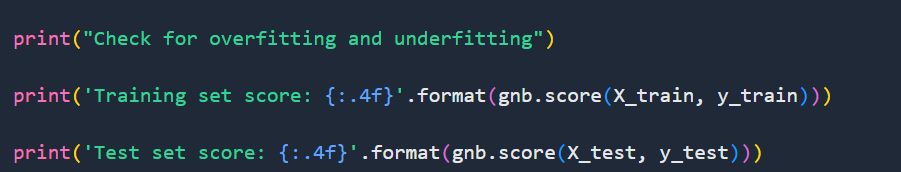
شکل ۲۰-۳. نمایش دقت پیش بینی برای داده های آزمون و آموزش در پایتون

خروجی کد شکل ۲۰-۳ را در شکل ۲۱-۳ می‌توانید ببینید.



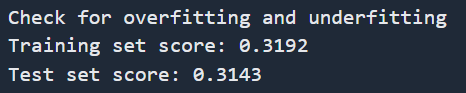
شکل ۲۱-۳. خروجی کد شکل ۲۰-۳.

علاوه بر این ها در یادگیری ماشین، ما معمولاً با دو مفهوم اصلی overfitting و کم‌شدن underfitting روبرو هستیم. این دو مفهوم نشان دهنده این هستند که مدل به چه میزان با داده‌های آموزش سازگار است. زیاده روی در یاددهی (overfitting) به موقعیتی گفته می‌شود که مدل به گونه‌ای بر روی داده‌های آموزش بسیار دقیق است که تقریباً به خاطر بسیاری از جزئیات ناهمخوانی در داده‌های جدید به خوبی عمل نمی‌کند. برای مثال، اگر یک مدل به صورت بسیار دقیقی بر روی داده‌های آموزش عمل کند، اما بر روی داده‌های تست نتایج ناامید کننده‌ای داشته باشد، به این مفهوم گفته می‌شود که مدل اضافه‌شده است. کم‌شدن (underfitting) به موقعیتی گفته می‌شود که مدل به گونه‌ای نسبتاً خیلی ساده است که با داده‌های آموزش به درستی کار نمی‌کند و در نتیجه نتایج بدی را در همه داده‌ها (آموزش و تست) تولید می‌کند. برای مثال، اگر یک مدل به صورت بسیار ساده‌ای طراحی شود، به گونه‌ای که بر روی داده‌های آموزش هم به درستی کار نکند در نتیجه بر روی داده‌های تست نیز نتایج بدی تولید می‌کند، به این مفهوم گفته می‌شود که مدل کم‌شده است.



شکل ۲۲-۳. چک کردن overfitting و underfitting

خروجی کد شکل ۲۲-۳:



شکل ۲۳-۳. خروجی کد شکل ۲۲-۳.

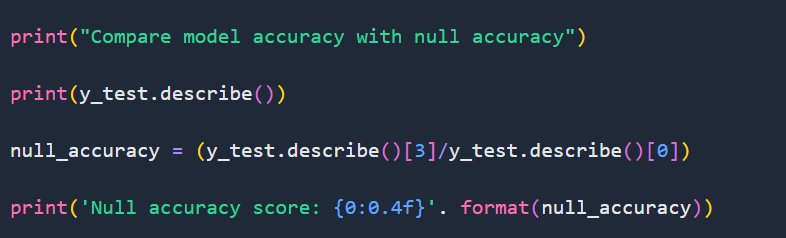
با توجه به شکل ۲۳-۳ خروجی overfitting و underfittingنداریم.

### ۲-۷-۳ معیار دقت صفر (Null Accuracy)

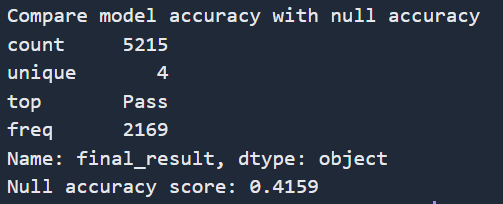
علاوه بر این معیارها، معیار دقت صفر (Null Accuracy) نیز برای ارزیابی مدل‌های دسته‌بندی مورد استفاده قرار می‌گیرد دقت صفر ، نسبت تعداد نمونه‌هایی است که با تعلیم مدل، برچسب غیرفعال (برچسب اکثریت کلاس) دریافت می‌کنند به کل تعداد نمونه‌ها. به عنوان مثال، اگر مدل دسته‌بندی با دو کلاس A و B داشته باشیم و تعداد نمونه‌های کلاس A ۴۰۰ و تعداد نمونه‌های کلاس B ۱۰۰ باشد، دقت صفر برابر با ۸۰٪ خواهد بود. به عبارت دیگر، اگر مدلی با دقت کمتر از ۸۰٪ داشته باشیم، می‌توانیم بگوییم که مدل بهتر از یک مدل که همیشه برچسب اکثریت کلاس اکثریت را پیش‌بینی می‌کند، نیست.

معیار دقت صفر در برخی موارد می‌تواند مفید باشد، به عنوان مثال در مواردی که کلاس‌ها نامتوازن باشند و تعداد نمونه‌های یک کلاس بسیار بیشتر از نمونه‌های دیگر باشد. در این موارد، مدلی که همیشه برچسب اکثریت را پیش‌بینی کند، می‌تواند دقت بالایی داشته باشد، اما به علت پیش‌بینی نامتوازن، نمی‌تواند به عنوان یک مدل خوب شناخته شود. در این شرایط، معیار دقت صفر می‌تواند به عنوان یک معیار مقایسه‌ای مفید برای این‌که ببینیم مدل ما به چه میزان بهتر از پیش‌بینی اکثریت عمل می‌کند، مورد استفاده قرار گیرد.

در تکه کد پایتون شکل ۲۴-۳ دقت صفر را محاسبه می کنیم.



شکل ۲۴-۳. محاسبه دقت صفر.



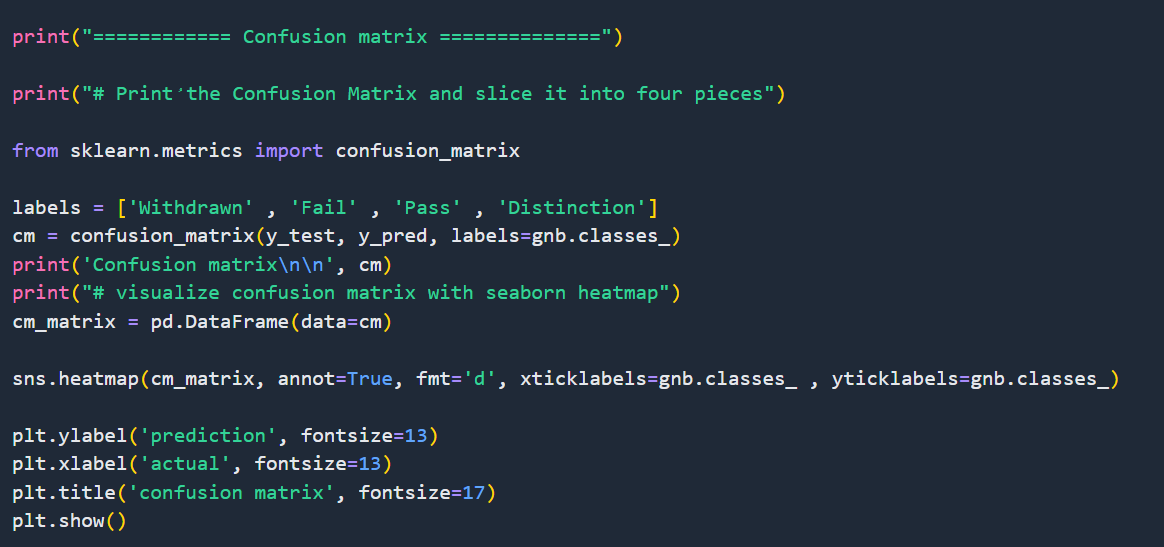
شکل ۲۵-۳. خروجی کد شکل ۲۴-۳.

با مقایسه دقت صفر و دقت (accuracy) متوجه می شویم دقت این پیش بینی پایین تر از دقت صفر بوده و مطلوب نیست.

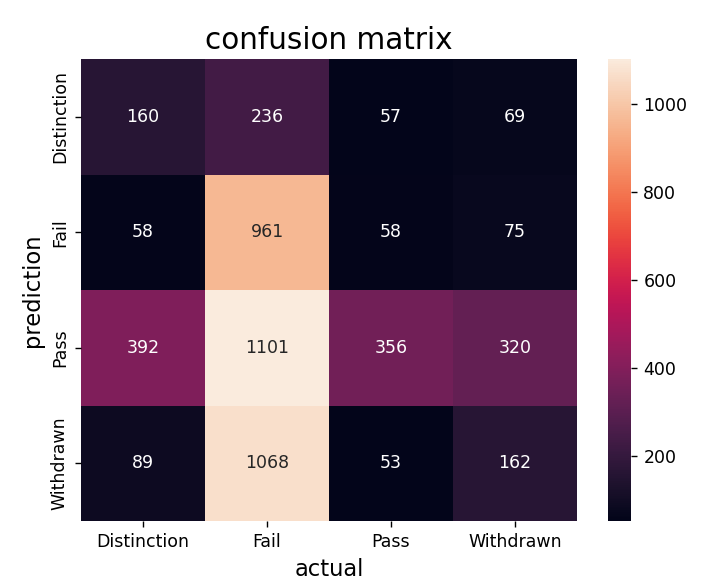
### ۳-۷-۳ ماتریس اغتشاش

ماتریس اغتشاش در این مسئله دسته‌بندی که چهار نتیجه ممکن دارد (fail,pass ,distinction , withdrawn)، در یک جدول نمایش داده می شود. ردیف‌های جدول به کلاس‌های واقعی اشاره دارند و ستون‌ها به کلاس‌های پیش‌بینی شده مربوط می‌شوند. در این ماتریس اغتشاش، قطر اصلی (از بالا سمت چپ تا پایین سمت راست) تعداد پیش‌بینی‌های صحیح را برای هر کلاس نشان می‌دهد، در حالی که عناصر خارج از قطر، میزان اشتباهات پیش‌بینی را نشان می‌دهند.

تکه کد پایتون زیر ماتریس اغتشاش این پیش بینی را به دست آورده و نشان می دهد:



شکل ۲۶-۳. نمایش confusion matrix در پایتون



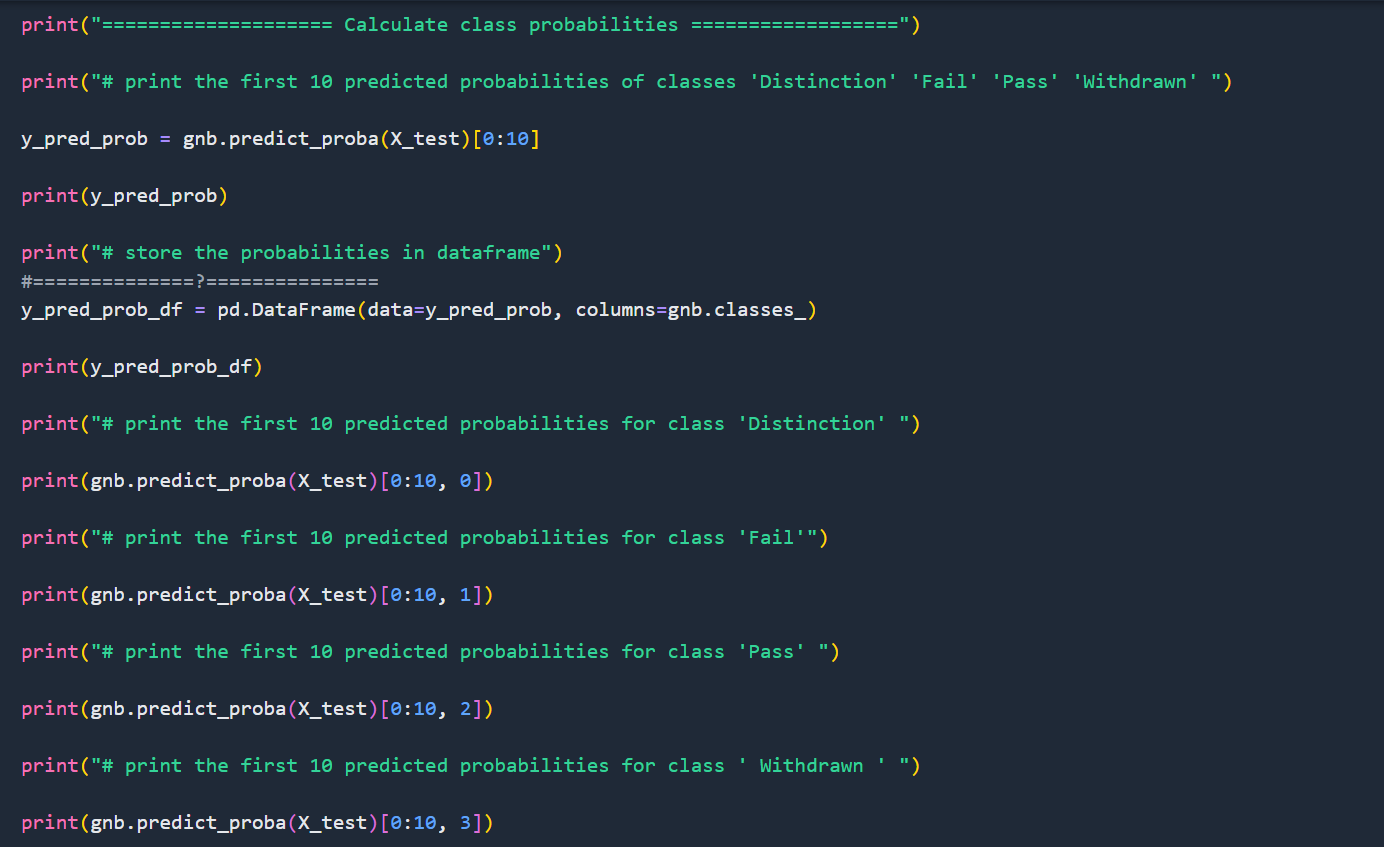
شکل ۲۷-۳. خروجی کد شکل ۲۶-۳. سطر ها نشان دهنده پیش‌بینی برچسب و ستون ها نشان دهنده برچسب واقعی است.

### ۴-۷-۳ احتمالات چند کلاسه

محاسبه احتمالات چند کلاسه (multi classes probabilities) جزو معیارهای ارزیابی پیش‌بینی مدل در ماشین لرنینگ است. این معیار به عنوان یکی از روش‌های ارزیابی کیفیت مدل برای دسته‌بندی چند کلاسه استفاده می‌شوند. با محاسبه احتمالات چند کلاسه، می‌توانیم ببینیم که مدل به چه اندازه موفق بوده است تا دسته بندی صحیح را برای داده های ورودی ارايه دهد.نمودار هیستوگرام (histogram)می‌تواند به عنوان یک ابزار مفید برای بررسی کیفیت مدل در پیش‌بینی چند کلاسه استفاده شود. با توجه به شکل و توزیع احتمالات چند کلاسه روی نمودار هیستوگرام، می‌توانیم بررسی کنیم که آیا مدل به درستی دسته‌بندی‌های خود را انجام داده است یا خیر. برای بررسی کیفیت مدل با استفاده از نمودار هیستوگرام به اینصورت است:

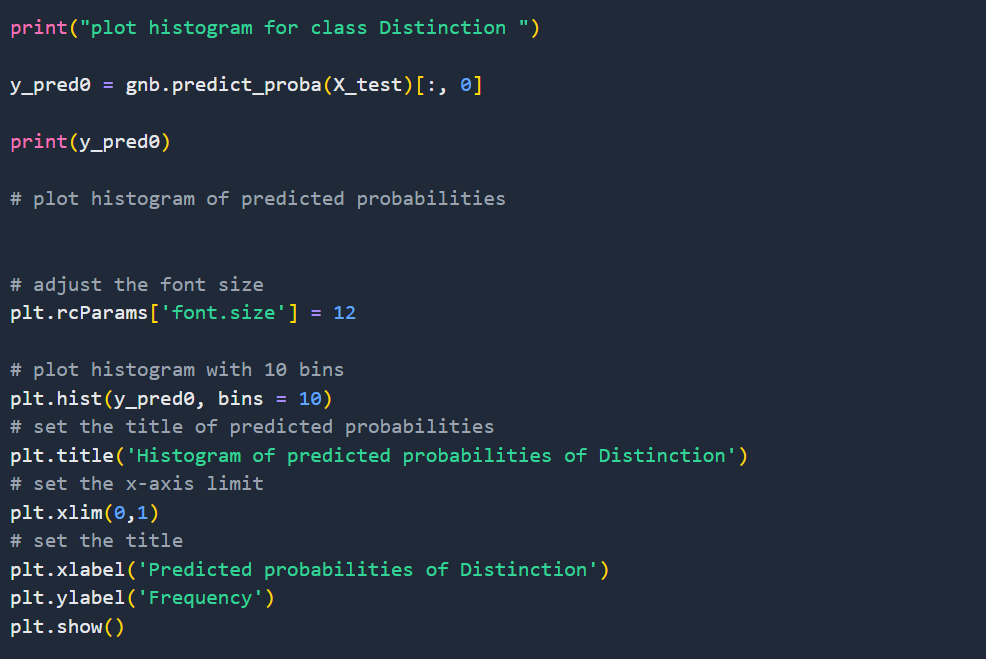
* بررسی توزیع احتمالات: با بررسی توزیع احتمالات چند کلاسه روی نمودار هیستوگرام، می‌توانیم بررسی کنیم که آیا مدل به درستی دسته‌بندی‌های خود را انجام داده است یا خیر. برای مثال، اگر توزیع احتمالات برای هر کلاس بسیار مشابه باشد و همه کلاس‌ها به یک شکل بر روی نمودار هیستوگرام قرار داشته باشند، این نشان دهنده یک مدل خوب است که به درستی دسته‌بندی می‌کند. اما اگر توزیع احتمالات بسیار متفاوت باشد و برخی کلاس‌ها به شدت بیش‌ بار(Overfit) شده باشند، این نشان دهنده وجود مشکل در دسته‌بندی مدل است.
* بررسی تفاوت بین احتمالات: با بررسی تفاوت بین احتمالات چند کلاسه روی نمودار هیستوگرام، می‌توانیم بررسی کنیم که آیا مدل به درستی دسته‌بندی‌های خود را انجام داده است یا خیر. برای مثال، اگر بین احتمالات کلاس‌ها تفاوت زیادی وجود داشته باشد، این نشان دهنده وجود مشکل در دسته‌بندی مدل است.
* بررسی انحراف از مقدار مورد انتظار: با بررسی انحراف از مقدار مورد انتظار برای هر کلاس، می‌توانیم بررسی کنیم که آیا مدل به درستی دسته‌بندی‌های خود را انجام داده است یا خیر. برای مثال، اگر احتمالات مدل برای یک کلاس به طور قابل توجهی از مقدار مورد انتظار برای آن کلاس کمتر باشد، این نشان دهنده وجود مشکل در دسته‌بندی مدل است.

با توجه به موارد فوق، می‌توان نمودار هیستوگرام رابرای بررسی کیفیت مدل در پیش‌بینی چند کلاسه به کار برد. تکه کد شکل ۳-۲۸ احتمال پیش بینی ۴ برچسب fail , pass , distinction ,withdrawn را ابتدا برای ۱۰ نمونه از داده ی آزمون و سپس جداگانه برای هر برچسب نشان می دهد.

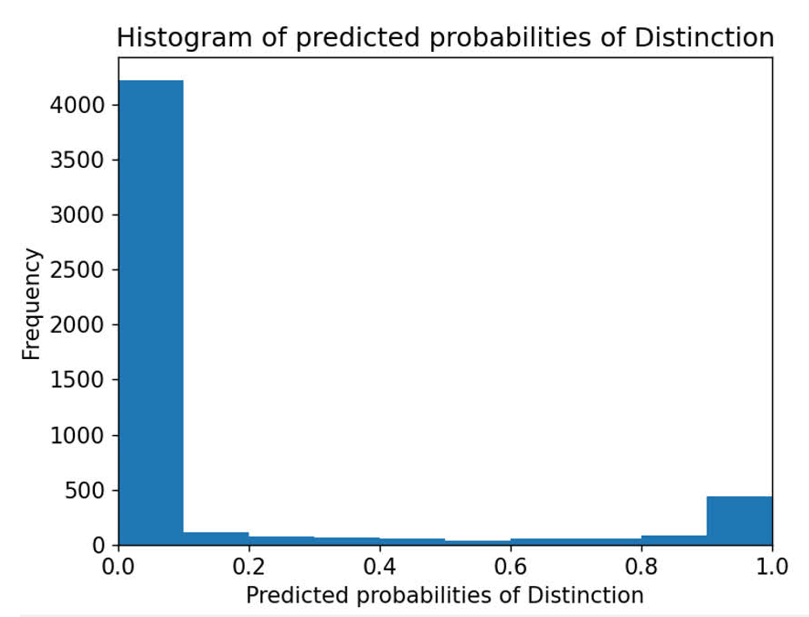


شکل ۳-۲۸. نمایش احتمالات پیش‌بینی ۱۰ نمونه.

تکه کد شکل ۳-۲۹، نمودار هیستوگرام برای احتمالات پیش‌بینی کلاس Distinction را نمایش می‌دهد که خروجی آن در شکل ۳۰-۳ نمایش داده شده است.

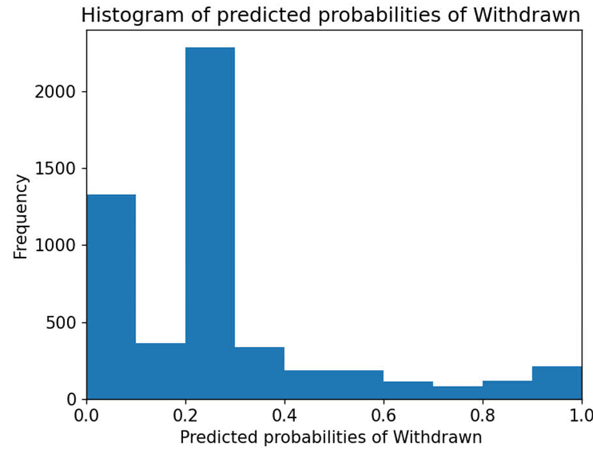
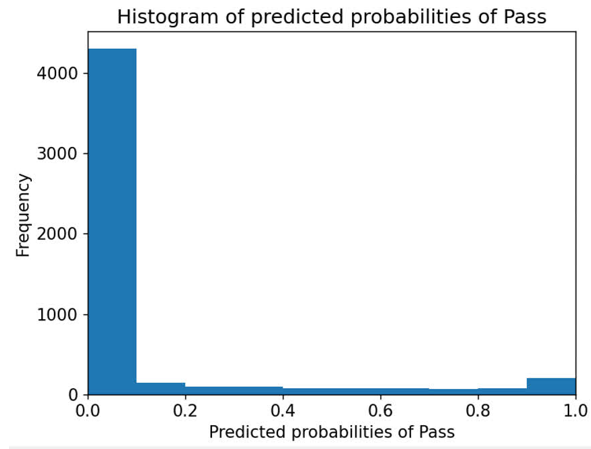
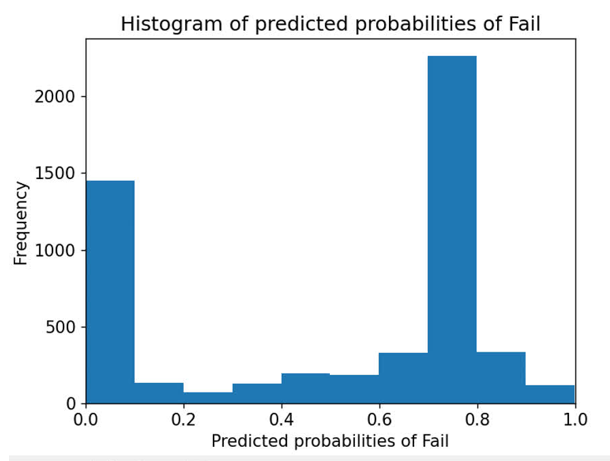


شکل۲۹ -۳ کد نمایش نمودار هیستوگرام برای احتمالات پیش بینی کلاس Distinction



شکل۳۰ -۳ نمایش گرافیکی نمودار هیستوگرام Distinction

در شکل ۳۱-۳ بقیه نمودار هیستوگرام را برای بقیه برچسب ها رسم شده. همانطور که در شکل ۳۱-۳ می‌بینیم تفاوت نمودار های هیستوگرام زیاد است و این نشان از دسته بندی و پیشبینی نامطلوب دارد.



شکل ۳۱-۳. به ترتیب از راست به چپ نمودار هیستوگرام برچسب های Fail و Pass و Withdrawn.

### ۴-۷-۳ معیار precision

معیار precision یکی از معیارهای ارزیابی عملکرد یک سیستم هوش مصنوعی است که به طور کلی برای ارزیابی دقت و صحت پاسخ‌های سیستم استفاده می‌شود. این معیار، نسبت تعداد پاسخ‌های درست به تعداد کل پاسخ‌های داده‌شده توسط سیستم را نشان می‌دهد. فرمول کلی این معیار به صورت زیر است:

ماتریس اغتشاشی که ما داشتیم ، یک ماتریس اغتشاش بود. معیار precision برای هر برچسب اینطور محاسبه می شود:

# فصل 4 نتیجه‌گیری

## 4-1 مقدمه

# مراجع

در این قسمت، تمامی مراجع استفاده شده در متن فهرست می‌شوند. تمام مراجع مندرج در این بخش، باید حداقل یک‌بار در متن اصلی استفاده شده باشند. تمام مراجع لارم است با فرمت یکسان مانند مثال زیر نمایش داده شوند.

[1] Smith J., Jones A., Andersen P., Jameson G., “Here is the title of this particular journal paper,” *Journal Name*, 31(2), 105–112, 2015.

شیوه‌ی ارائه مراجع فارسی و انگلیسی

1. نام خانوادگی، نام (مولفان و مترجمان)؛ *عنوان اصلی كتاب*؛ عنوان فرعی كتاب (جزئیات عنوان كتاب در صورت وجود داخل پرانتز)، نام سایر افراد دخیل در تالیف یا ترجمه، ناشر، محل انتشار، شماره جلد، شماره ویرایش، سال انتشار به عدد‌.‌
2. منهاج، محمد‌باقر؛ *هوش محاسباتی* (جلد اول: مبانی شبكه‌های عصبی)، انتشارات دانشگاه صنعتی امیر‌كبیر، تهران، ویرایش اول، 1379‌.‌
3. نام خانوادگی، نام؛ نام‌‌خانوادگی و نام مولف دوم؛ مولف سوم؛ ”عنوان مقاله به‌صورت عادی و داخل گیومه“، *نام كامل مجله به صورت ایتالیك*، شماره دوره یا جلد، شماره مجله، شماره صفحات، سال انتشار‌.‌
4. نام خانوادگی، نام مجری؛ *عنوان طرح پژوهشی به‌صورت ایتالیك*، شماره ثبت، نام كامل محل انجام و سفارش دهنده، سال انجام طرح‌.‌
5. نام شركت/نام فرد؛ عنوان صفحه؛ *آدرس اینترنتی*.
6. Book authors’ names; *Book Title in Italic* (and the title components, if any), Edition number, Publisher, Date of publish.
7. Van de Vegte, J.; *Feedback Control Systems*, 2nd Edition, Prentice Hall, 1990
8. Authors’ names separated by comma-dots; “The paper title in Regular Times New Roman 12pt”, *Paper Address in Italic*, Publishing Place, paper page, Year of Publish.
9. Safonov, M.; “Stability margins of diagonally perturbed multivariable feedback systems”, *IEEE Proceedings*, Part D, p. p. 251-256, Nov. 1982.
10. Company Name/ Person Name; Page Title; *Internet Address*.

واژه نامه انگلیسی به فارسی

در شرایطی که اصطلاحات مورد استفاده نو یا در ادبیات فنی، ناآشنا باشد، تدوین واژه‌نامه برای پایان‌نامه/رساله توصیه می‌شود. لازم به ذکراست تهیه این واژه‌نامه اختیاری بوده و ضرورت تهیه آن را استاد راهنما تعیین می‌نماید.

تهیه متن واژه‌نامه به شرح زیر توصیه می‌شود:

A

|  |  |
| --- | --- |
| دانشگاهی | Academic |
| تحلیلی | Analytical |

پیوست‌ها

در صورت لازم، این قسمت ذکر میشود و در عیر این صورت آن را حذف کنید. این قسمت مطالبی که مرتبط با متن پایان نامه بوده اما ذکر آنها در متن اصلی موجب خستگی خواننده یا سردرگمی او شود، گنجانده می شود.

پیوست الف: عنوان پیوست

جدول ۲-۱ به صورت نمونه پیشنهاد شده است.

جدول ‏0‑1: نتایج حاصل از سنتز دو مدار

|  |  |  |  |
| --- | --- | --- | --- |
| Power  () | Area  () | Delay  () | n |
| 556.20 | 1716.00 | 1.90 | N=8 |
| 1200.01 | 3641.00 | 2.70 | N=16 |

**Abstract**

In this part, the abstract of the thesis must be presented, including supposes, aims and results. In this part, the abstract of the thesis must be presented, including supposes, aims and results. In this part, the abstract of the thesis must be presented, including supposes, aims and results. In this part, the abstract of the thesis must be presented, including supposes, aims and results.

**Keywords:** …, …, …, …, …



Shahid Rajaee Teacher Training University

Faculty of Computer Engineering

Department of……

B. Sc. Thesis

Title:

**………………….**

Supervisor:

Dr. …………..

By:

………………….

Winter 2020