مبانی علم داده

پروژه پایانی

آدرینا ابراهیمی

۹۹۳۶۲۳۰۰۲

کیان مجلسی

۹۹۳۶۱۳۰۵۱

دکتر رضاپور

خرداد ۱۴۰۳

فهرست مطالب

مقدمه

در دنیای امروز، علم داده و یادگیری ماشین به عنوان ابزارهای قدرتمندی برای تحلیل و استخراج دانش از داده‌ها به کار گرفته می‌شوند. این پروژه با هدف استفاده از مدل‌های مختلف یادگیری ماشین برای حل دو مسئله مهم و چالش‌برانگیز در حوزه سلامت انجام می‌شود: رگرسیون و دسته‌بندی. در مسئله رگرسیون، از مجموعه‌داده تومور مغزی استفاده خواهد شد. این مجموعه‌داده شامل اطلاعاتی از بیماران مبتلا به تومور مغزی است و هدف اصلی مسئله پیش‌بینی تعداد ماه‌های زنده ماندن این بیماران با توجه به وضعیت‌های مختلف بالینی و داده‌های پزشکی آنهاست. پیش‌بینی دقیق مدت زمان زنده ماندن می‌تواند به پزشکان در برنامه‌ریزی درمان‌ها و ارائه مشاوره‌های دقیق‌تر به بیماران کمک کند. در مسئله دسته‌بندی، از مجموعه‌داده کوید۱۹ استفاده خواهد شد. این مجموعه‌داده شامل داده‌هایی از بیماران مبتلا به کووید-۱۹ است و هدف دسته‌بندی این بیماران به دو گروه "زنده مانده" و "زنده نمانده" است. این دسته‌بندی می‌تواند به پیش‌بینی نتایج بیماری و تصمیم‌گیری‌های سریع‌تر و مؤثرتر در مدیریت بیماران کمک کند. هدف نهایی این پروژه ارائه و مقایسه مدل‌هایی با دقت و کارایی بالا است که بتوانند به طور مؤثری در حل این دو مسئله مهم به کار گرفته شوند.

مسئله رگرسیون: پیش‌بینی تعداد ماه‌های زنده ماندن بیماران دارای تومور مغزی

رگرسیون یکی از تکنیک‌های مهم در یادگیری ماشین و آمار است که برای مدل‌سازی و تحلیل رابطه بین یک متغیر وابسته (یا خروجی) و یک یا چند متغیر مستقل (یا ورودی) به کار می‌رود. هدف اصلی رگرسیون پیش‌بینی مقدار متغیر وابسته بر اساس مقادیر متغیرهای مستقل است. رگرسیون انواع مختلفی دارد که متناسب با نوع و پیچیدگی مسئله می‌توان مدل مناسب را انتخاب کرد. تومور مغزی نوعی توده یا رشد غیرطبیعی سلولی در مغز یا نواحی نزدیک به آن است. تومورهای مغزی می‌توانند خوش‌خیم (غیرسرطانی) یا بدخیم (سرطانی) باشند. این تومورها ممکن است از سلول‌های مغزی، غشاهای اطراف مغز، اعصاب جمجمه‌ای یا از سایر قسمت‌های بدن که به مغز سرایت کرده‌اند، تشکیل شوند. علائم تومور مغزی بسته به نوع، اندازه و محل تومور متفاوت است و ممکن است شامل سردرد، مشکلات بینایی، تشنج و تغییرات شناختی یا شخصیتی باشد. تشخیص و درمان به موقع می‌تواند تأثیر بسزایی در بهبود وضعیت بیمار داشته باشد. مسئله رگرسیون در این پروژه به پیش‌بینی تعداد ماه‌های زنده ماندن بیماران مبتلا به تومور مغزی اختصاص دارد. با استفاده از داده‌های بالینی و پزشکی بیماران، مدل‌های یادگیری ماشین تلاش می‌کنند تا رابطه بین ویژگی‌های مختلف بیمار و مدت زمان زنده ماندن او را بیابند.

مجموعه‌داده

این مجموعه داده شامل اطلاعات مربوط به بیماران مبتلا به تومور مغزی و الگوهای بازگشت تومور بر اساس مراحل مختلف است. این مجموعه‌داده از ۱۱ ویژگی منحصر به فرد و ۲۰۰۰ بیمار یکتا تشکیل شده است. NaN نشان‌دهنده داده‌های مفقود است. ستون‌های موجود در این مجموعه داده عبارتند از:

- PatientID: شناسه منحصر به فرد هر بیمار شامل مقادیر عددی
- Age: سن بیمار شامل مقادیر عددی
- Gender: جنسیت بیمار شامل دو مقدار Male و Female
- Tumor Type: نوع تومور شامل مقادیر Meningioma، Astrocytoma و Glioblastoma
- Tumor Grade: مرحله پیش‌روی تومور شامل مقادیر I، IV، II و III
- Tumor Location: محل قرارگیری تومور در مغز شامل مقادیر Frontal lobe، Parietal lobe، Temporal lobe و Occipital lobe
- Treatment: نوع درمان شامل مقادیر زیر
  - Surgery + Radiation
  - Surgery + Chemotherapy
  - Surgery
  - Chemotherapy
  - Radiation
  - Surgery + Radiation therapy
  - Chemotherapy + Radiation
- Treatment Outcome: نتیجه درمان شامل مقادیر Progressive disease، Complete response، Stable disease و Partial response
- Time to Recurrence (months): مدت زمان بازگشت تومور به ماه شامل مقادیر عددی
- Recurrence Site: محل قرارگیری تومور در مغز پس از بازگشت شامل مقادیر Frontal lobe، Parietal lobe، Temporal lobe و Occipital lobe
- Survival Time (months): تعداد ماه‌های زنده ماندن پس از تشخیص شامل مقادیر عددی

	Patient ID	Age	Gender	Tumor Type	Tumor Grade	Tumor Location	Treatment	Treatment Outcome	Time to Recurrence (months)	Recurrence Site	Survival Time (months)
0	1	45	Male	Glioblastoma	IV	Frontal lobe	Surgery	Partial response	10.0	Temporal lobe	18
1	2	55	Female	Meningioma	I	Parietal lobe	Surgery	Complete response	NaN	NaN	36
2	3	60	Male	Astrocytoma	III	Occipital lobe	Surgery + Chemotherapy	Progressive disease	14.0	Frontal lobe	22
3	4	50	Female	Glioblastoma	IV	Temporal lobe	Surgery + Radiation therapy	Complete response	NaN	NaN	12
4	5	65	Male	Astrocytoma	II	Frontal lobe	Surgery + Radiation therapy	Partial response	24.0	Frontal lobe	48

بینش دریافتی از مجموعه داده

قبل از پیاده‌سازی هر گونه مدل یادگیری ماشین با تحلیل مجموعه داده، می‌توانیم به بینش‌های زیر دست یابیم:

۱. **توزیع داده‌ها** :شناسایی الگوها و توزیع ویژگی‌ها مانند سن، جنسیت و مرحله تومور.

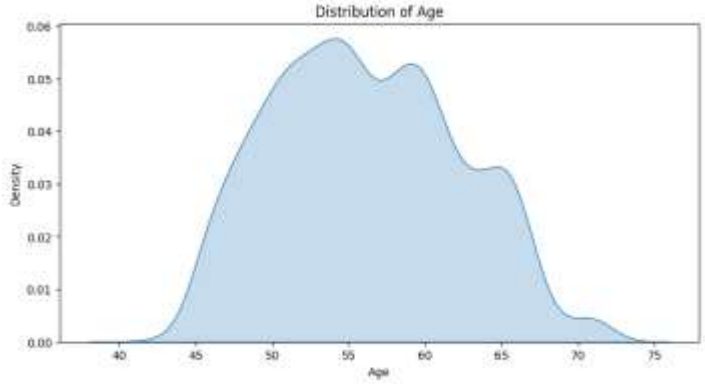
۲. **ارتباط ویژگی‌ها** :بررسی همبستگی بین متغیرها برای شناسایی ویژگی‌های مهم و تأثیرگذار.

این تحلیل‌ها پایه‌ای قوی برای ساخت مدل‌های دقیق‌تر و مؤثرتر فراهم می‌کنند.

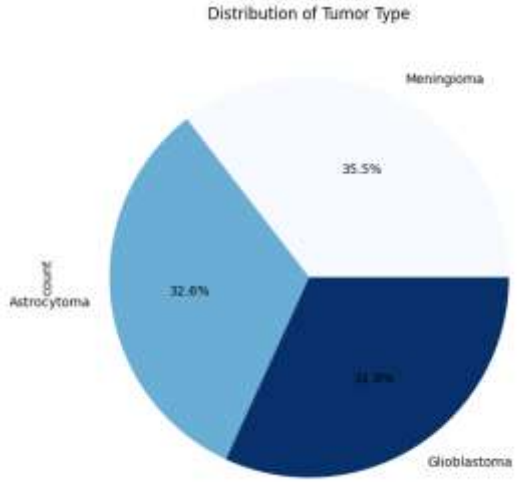
توزیع داده‌ها

در این قسمت به بررسی توزیع داده‌ها با توجه به ویژگی‌های مختلف می‌پردازیم.

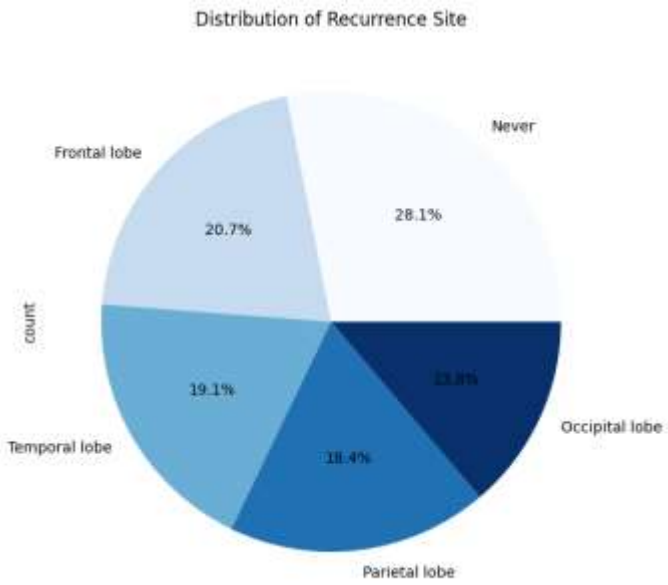
در ابتدا با بررسی نمودار زیر به توزیع سنی بیماران دارای تومور مغزی پی می‌بریم. مشاهده می‌شود سن بیشتر بیماران بین ۵۰ تا ۵۵ خواهد بود و در مرحله بعد اکثر بیماران دارای ۶۰ سال سن هستند.



حال با مشاهده نمودار درصد نوع هر تومور، پی می‌بریم انواع تومورهای مغزی موجود در این مجموعه‌داده دارای سهم تا حدودی یکسان هستند.



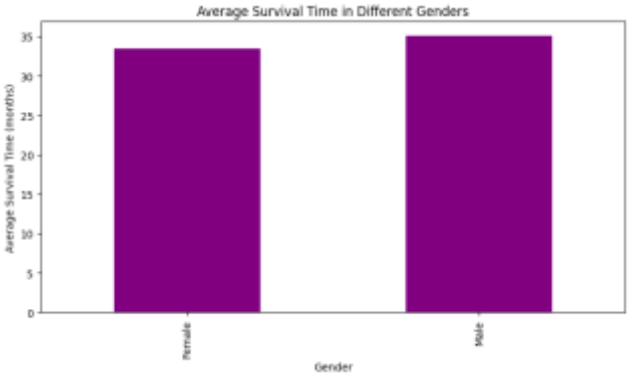
اکنون، با رسم نمودار ناحیه قرارگیری تومور در صورت بازگشت بیماری، به این نکته پی می‌بریم که ۲۸ درصد از بیماران پس از درمان مجددا درگیر این بیماری نشده‌اند. (با فرض اینکه در صورت بازگشت تومور بیمار حتما به بیمارستان مراجعه کرده و اطلاعاتش ذخیره می‌شود.) در صورت بازگشت تومور بیشترین ناحیه‌ای که تومور مجددا در آن مشاهده می‌شود قسمت Frontal Lobe مغز بوده و در درصد کمتری از بیماران پس از درمان، تومور در قسمت Occipital Lobe مشاهده شده است.



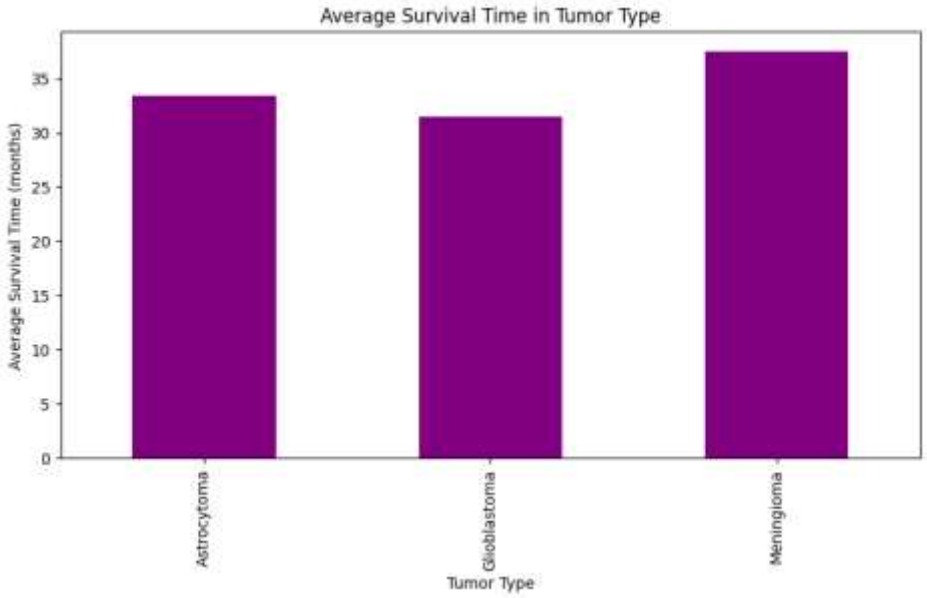
ارتباط ویژگی‌ها

در این قسمت به بررسی ارتباط بین ویژگی‌های مختلف می‌پردازیم.

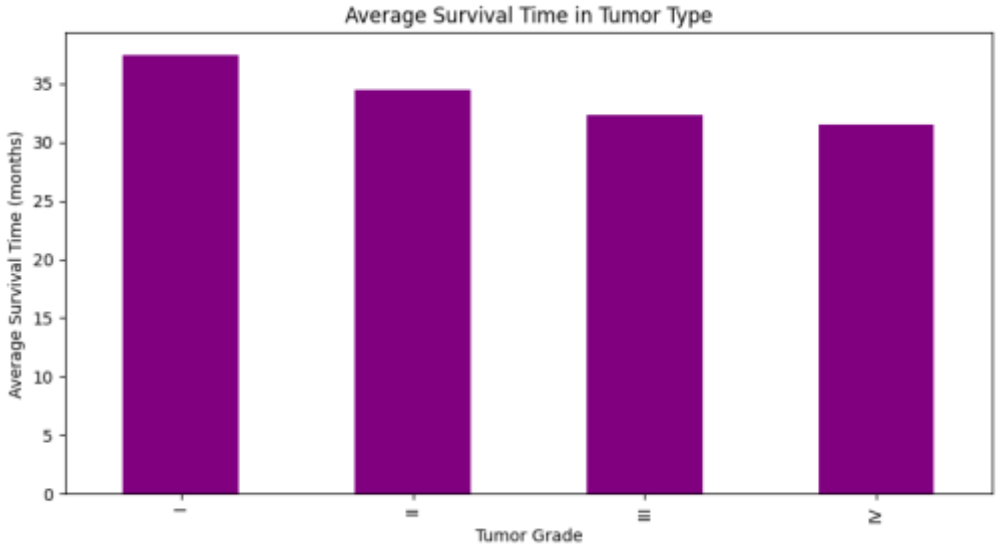
با رسم نمودار ارتباطی بین جنسیت و تعداد ماه‌های زنده ماندن در می‌یابیم جنسیت تاثیر چندانی در افزایش یا کاهش تعداد ماه‌های زنده ماندن بیمار ندارد.



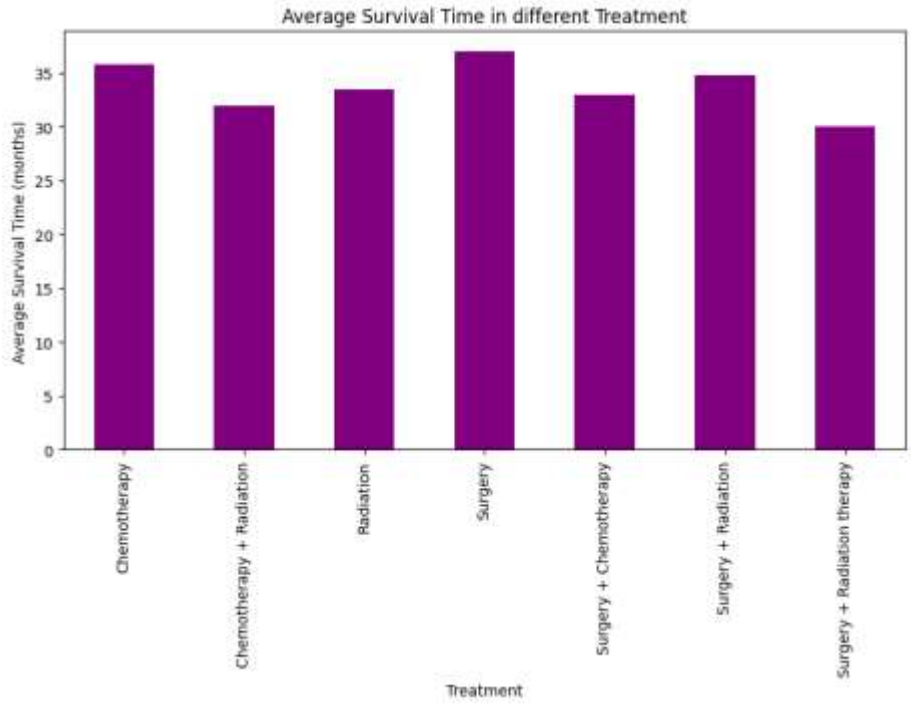
سپس می‌خواهیم ارتباط بین نوع تومور و میزان زنده ماندن بیمار را بررسی کنیم. با توجه به نمودار زیر، به طور میانگین افراد پس از تشخیص بیماری بین ۳۲ تا ۳۵ ماه زنده خواهند ماند. از بین این اشخاص، کسانی که دارای نوع تومور Meningioma هستند تعداد ماه‌های بیشتری را نسبت به افراد دارای انواع تومورهای دیگر زنده هستند.



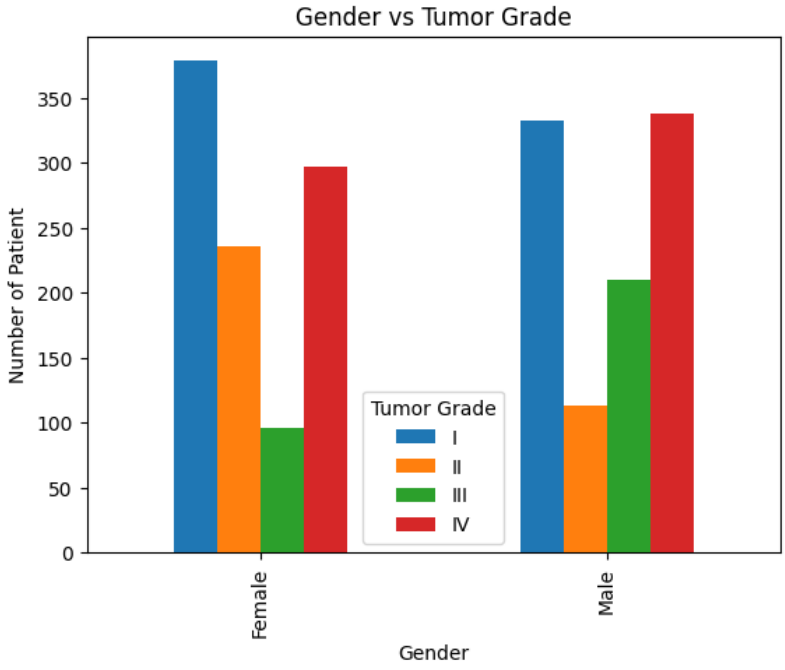
اکنون با رسم نمودار مرحله پیش‌روی تومور و ارتباط آن با میانگین میزان زنده بودن بیماران، همانطور که انتظار می‌رفت، مشاهده می‌کنیم هر چه تومور پیشرفته‌تر باشد، تعداد ماه‌های زنده ماندن افراد کمتر خواهد بود.



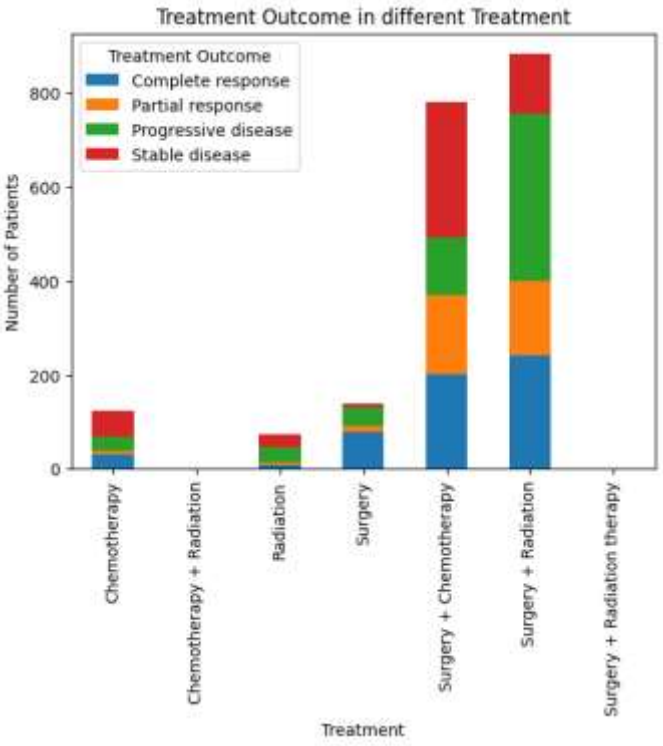
حال، می‌خواهیم بررسی کنیم هر یک از روش‌های درمان چه تاثیری در میانگین متغیر هدف داشته‌اند. مشاهده می‌شود در رتبه اول کسانی که جراحی شده‌اند و تومور به طور کامل برداشته شده‌است، تعداد ماه‌های بیشتری زنده مانده‌اند و در مرحله بعد کسانی که شیمی‌درمانی شده‌اند تعداد ماه‌های بیشتری را برای زندگی در اختیار داشته‌اند. بلعکس، کسانی که هم جراحی شده‌اند و هم پرتودرمانی شده‌اند، به طور میانگین ماه‌های کمتری را زنده بوده‌اند.



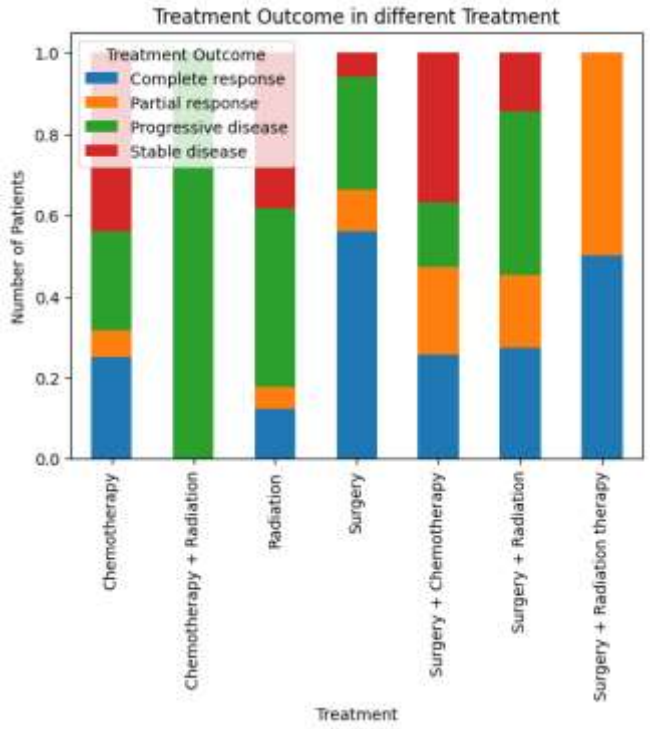
در این مرحله می‌خواهیم ارتباط بین جنسیت و سطوح مختلف پیش‌روی تومور را بررسی کنیم. مشاهده می‌شود تعداد زنانی که به تومور مغزی سطح اول مبتلا شده‌اند، بیشتر از مردان دارای تومور مغزی سطح اول هستند. بلعکس، تعداد مردانی که به پیشرفته‌ترین نوع تومور مغزی مبتلا شده بیشتر از زنان با همین سطح تومور مغزی هستند. در بین دو سطح میانی، زنان و مردان وضعیتی تقریبا مخالف یکدیگر دارند؛ تعداد زنان بیشتری نسبت به مردان به تومور سطح دوم مبتلا هستند و برعکس.



در مرحله بعد، می‌خواهیم بررسی کنیم هر روش درمان چه تاثیری روی بیماری اشخاص مورد بررسی داشته است. با نگاهی به نمودار در می‌یابیم افرادی که در کنار عمل جراحی شیمی‌درمانی شده یا هنگام عمل جراحی تومور پرتو دریافت کرده است، نتیجه بهتری از سایر افراد با روش‌های درمانی دیگر گرفته‌اند. (جواب کامل از درمان گرفته یا بیماری آن‌ها پایدار شده است.) چنین روشی را با در نظر گرفتن عوامل دیگر می‌توان به عنوان روش برتر درمان معرفی کرد. تعداد افرادی که هم زمان شیمی‌درمانی شده و پرتو دریافت کرده‌اند یا جراحی و پرتودرمانی شده‌اند، بسیار کم بوده برای همین نمی‌توان در مورد این روش درمان اظهار نظر کرد.

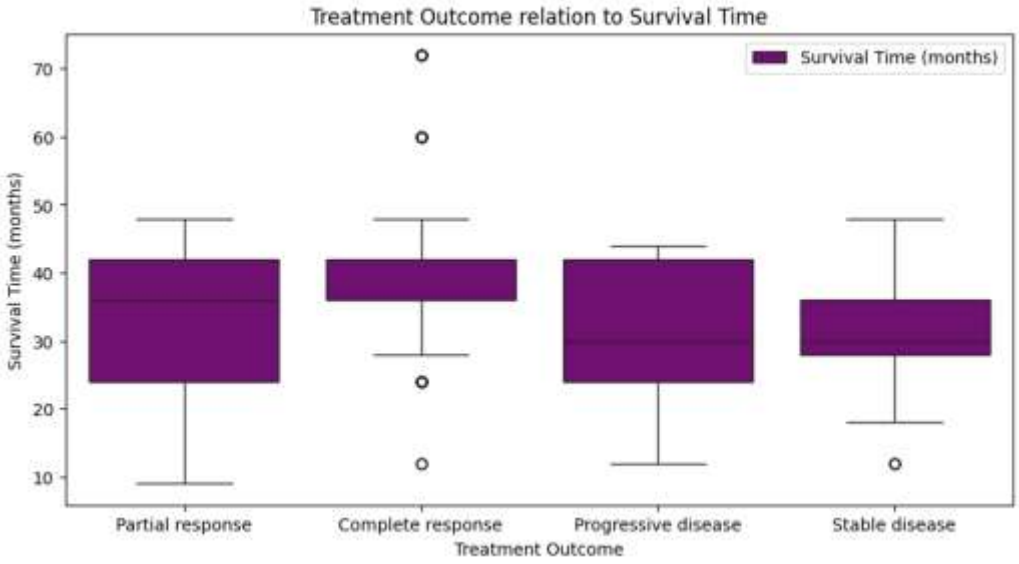


اگر نمودار قبل را نرمال‌سازی کنیم به نمودار زیر می‌رسیم. در این نمودار در صورتی که برای هر روش درمان تعداد زیادی داده وجود داشت، می‌توانستیم بهترین روش درمان را با توجه به نتایج کسب شده بیابیم. اما از آنجایی که برخی ستون‌ها مانند شیمی‌درمانی همراه با پرتو و جراحی همراه با پرتودرمانی تعداد داده‌های کمی دارند، نمی‌توان به طور کلی برای بهترین روش درمان نظری داد زیرا داده‌ها می‌توانند در شرایط حال سوگیری داشته باشند.



در آخر، می‌خواهیم مدت زمان زنده ماندن افراد را با توجه به نتیجه درمان آن‌ها بررسی کنیم. با توجه به نمودار جعبه‌ای زیر، مشاهده می‌کنیم همانطور که انتظار می‌رفت افرادی که نتیجه کامل از درمان گرفته‌اند ماه‌های بیش‌تری زنده بمانند و جعبه مربوط به این دسته

افراد بسیار محدودتر از سایرین است؛ بدین معنی که بیشتر این افراد بین ۳۵ الی ۴۲ ماه زنده می‌مانند. مشابه چنین وضعیتی تا حدودی برای افرادی که بیماری آن‌ها پس از درمان پایدار شده نیز مشاهده می‌شود؛ با این تفاوت که میانگین تعداد زنده ماندن این افراد حدود ۳۰ ماه است. اما در افرادی که بیماری پیش‌رونده دارند یا از درمان نتیجه کامل نگرفتند، چنین وضعیتی مشاهده نمی‌شود و تعداد ماه‌های زنده ماندن آن‌ها بازه وسیع‌تری نسب به اشخاص قبلی دارند.



پیش پردازش

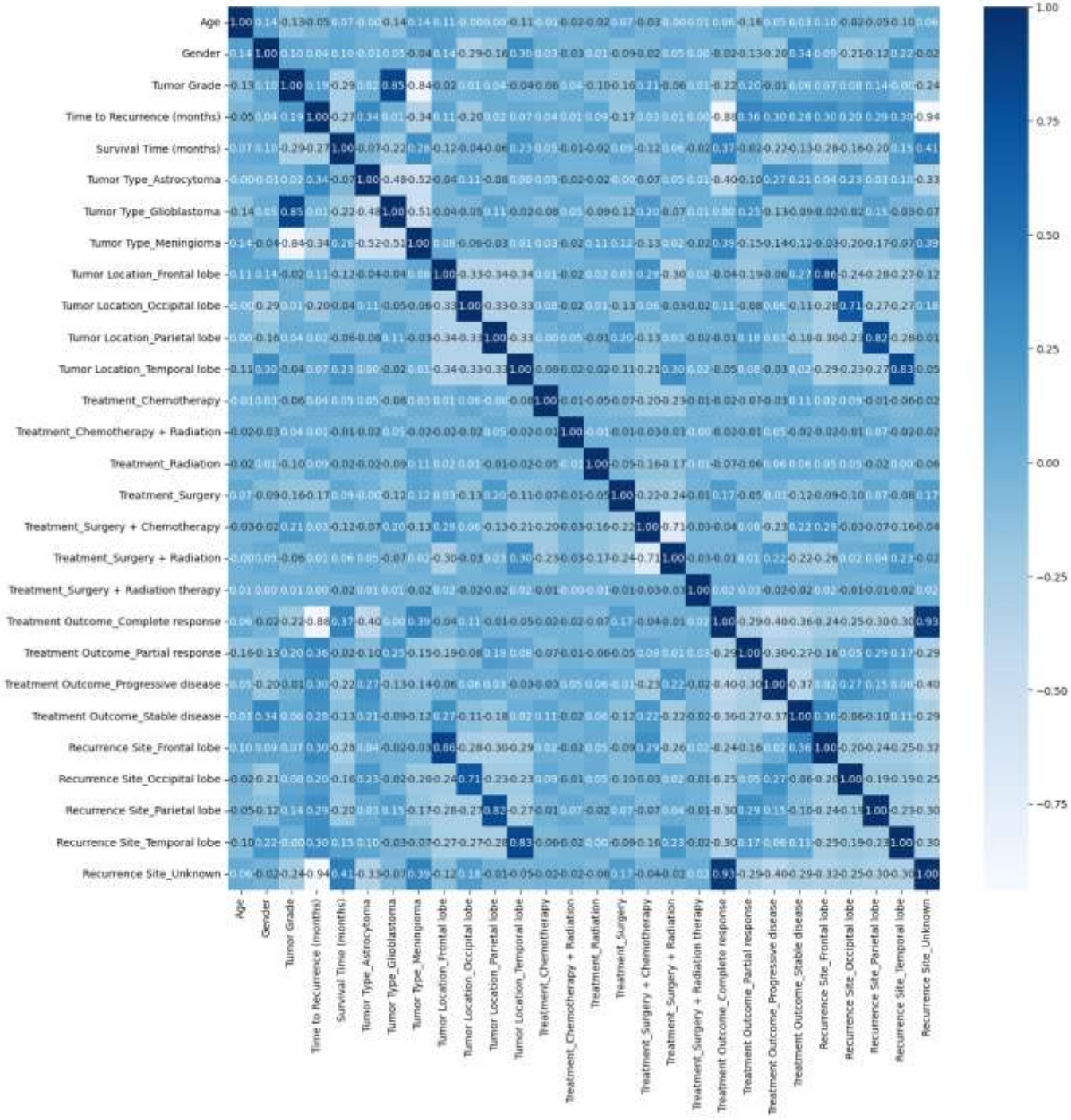
در ابتدا ستون PatientID شامل شناسه بیماران را حذف می‌کنیم. سپس، مقادیر خالی موجود در ستون Recurrence Site را با مقدار Never جایگزین می‌کنیم. (بدین معنی که بیمار مجددا دچار تومور مغزی نشده است در صورت بازگشت تومور حتما محل قرارگیری آن ثبت می‌شود.) حال، از آنجایی که ستون‌های Tumor Type، Tumor Location، Treatment، Treatment Outcome و Recurrence Site دارای مقادیر اسمی بدون ترتیب هستند و تعداد مقادیر متمایز آن زیادتر است، مقادیر آن‌ها را one-hot می‌کنیم و از آنجایی که ستون‌های Gender و Tumor Grade نیز دارای مقادیر اسمی هستند و مقادیر متمایز آن‌ها کمتر است یا دارای ترتیب هستند، از Label Encode استفاده می‌کنیم.

ماتریس همبستگی

ماتریس همبستگی جدولی مربعی است که در آن ضرایب همبستگی پیرسون بین هر دو متغیر در یک مجموعه داده نشان داده می‌شود. هرچه ضریب همبستگی بین دو متغیر به ۱ نزدیک‌تر باشد، نشان‌دهنده همبستگی قوی‌تر و مثبت‌تر بین آن دو است. به عبارت دیگر، مقادیر مثبت نشان می‌دهند که با افزایش یک متغیر، به احتمال زیاد متغیر دیگر نیز افزایش می‌یابد. در مقابل، مقادیر منفی نشان‌دهنده

همبستگی معکوس است، به این معنی که با افزایش یک متغیر، به احتمال زیاد متغیر دیگر کاهش می‌یابد. اکنون با توجه به این که تمامی مقادیر ستون‌های مجموعه‌داده دارای مقادیر عددی هستند، می‌توانیم ماتریس همبستگی بین ویژگی‌ها را رسم کنیم. با توجه به این ماتریس می‌توانیم نکات زیر را دریابیم:

- Tumor Grade: همبستگی منفی با زمان بقا دارد، یعنی کاهش درجه تومور ممکن است به افزایش زمان زنده ماندن منجر شود.
- Tumor Type_Astrocytoma: همبستگی مثبت نسبتاً قوی با زمان زنده ماندن دارد. این به این معناست که داشتن تومور از نوع Astrocytoma ممکن است با زمان بقای بیشتر همراه باشد.
- Tumor Type_Glioblastoma: همبستگی منفی قوی با زمان بقا دارد، یعنی داشتن تومور Glioblastoma ممکن است به کاهش زمان زنده ماندن منجر شود.
- Treatment_Chemotherapy + Radiation: همبستگی منفی با زمان زنده ماندن دارد، نشان‌دهنده این است که این نوع درمان ممکن است با زمان زنده ماندن کمتر همراه باشد.
- Treatment_Surgery: همبستگی مثبت با زمان زنده ماندن دارد، نشان‌دهنده این است که جراحی ممکن است با زمان زنده ماندن بیشتر همراه باشد.
- Treatment Outcome_Complete response: همبستگی مثبت نسبتاً قوی با زمان زنده ماندن دارد، یعنی گرفتن پاسخ کامل به درمان ممکن است به زمان زنده ماندن بیشتر منجر شود.
- Recurrence Site_Temporal lobe: همبستگی منفی قوی با زمان زنده ماندن دارد، نشان‌دهنده این است که وجود تومور در Temporal lobe ممکن است با زمان زنده ماندن کمتر همراه باشد.



در این مرحله، با بررسی مقادیر مینیمم، ماکسیمم و میانگین ستون‌های باقی‌مانده پی می‌بریم که می‌توانیم ستون‌های Age، Time to Recurrence (months) و Survival Time (months) را نرمال‌سازی کنیم. در مرحله آخر، ۸۰ درصد از داده‌ها را برای آموزش مدل‌ها و ۲۰ درصد از آن‌ها را برای تست مدل‌ها جداسازی می‌کنیم.

پیاده‌سازی مدل‌های یادگیری ماشین

در این بخش با استفاده از کتابخانه Scikit-learn در زبان پایتون به پیاده‌سازی انواع مدل‌های یادگیری ماشین از جمله رگرسیون خطی، درخت تصمیم، بردارهای پشتیبان، جنگل تصادفی، تقویت گرادیان، شبکه‌های عصبی چندلایه و k نزدیک‌ترین همسایه پرداخته و عملکرد هر یک از این مدل‌ها را با استفاده از معیارهای مختلف در مقایسه با یکدیگر بررسی می‌کنیم.

معیارهای ارزیابی

پس از آموزش مدل لازم است عملکرد آن را بسنجیم. این کار را می‌توان با استفاده از معیارهای مختلف مانند میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) انجام داد.

- **میانگین مربعات خطا (MSE)**: اختلاف بین مقادیر واقعی و پیش‌بینی‌شده توسط مدل.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **ریشه میانگین مربعات خطا (RMSE)**: جذر میانگین مربعات خطا، که برای بازگرداندن واحدها به مقیاس اصلی استفاده می‌شود.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **امتیاز R2**: نسبت واریانس توضیح داده شده توسط مدل به واریانس کل داده‌ها.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **میانگین قدرمطلق خطا (MAE)**: میانگین قدرمطلق اختلاف بین مقادیر واقعی و پیش‌بینی‌شده توسط مدل.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

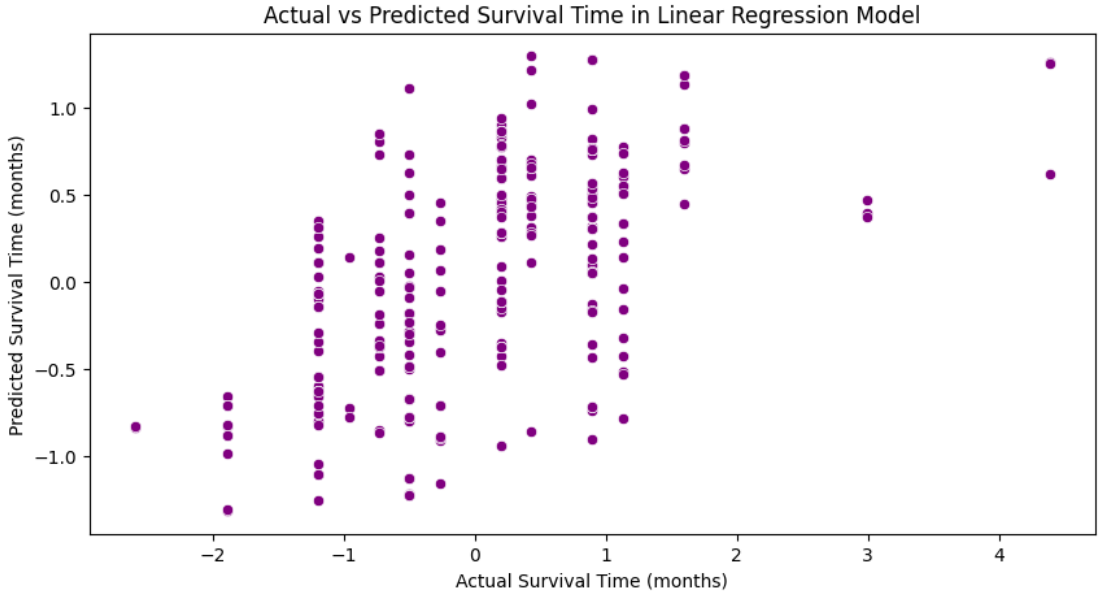
رگرسیون خطی

رگرسیون خطی یکی از ساده‌ترین و پرکاربردترین تکنیک‌های آماری و یادگیری ماشین است که برای مدل‌سازی و تحلیل رابطه بین یک متغیر وابسته و یک یا چند متغیر مستقل استفاده می‌شود. در این روش، فرض می‌شود که رابطه بین متغیرها خطی است و مدل به صورت یک خط راست بیان می‌شود. هدف یافتن خطی است که کمترین خطا را داشته باشد. با استفاده از ماژول رگرسیون خطی در کتابخانه

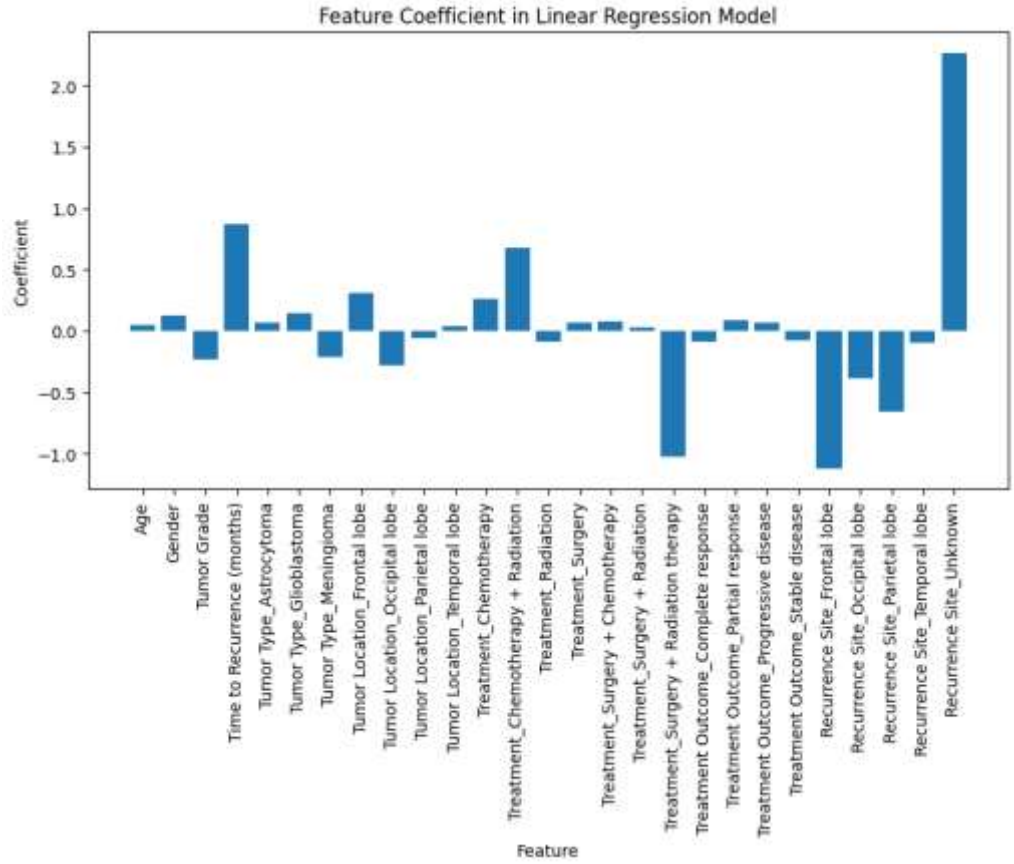
ذکر شده یک مدل رگرسیون خطی را آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل رگرسیون خطی معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.6411591803134792
Root Mean Squared Error: 0.8007241599411618
R2 Score: 0.37783223968880675
Mean Absolute Error: 0.6144308314507859
```

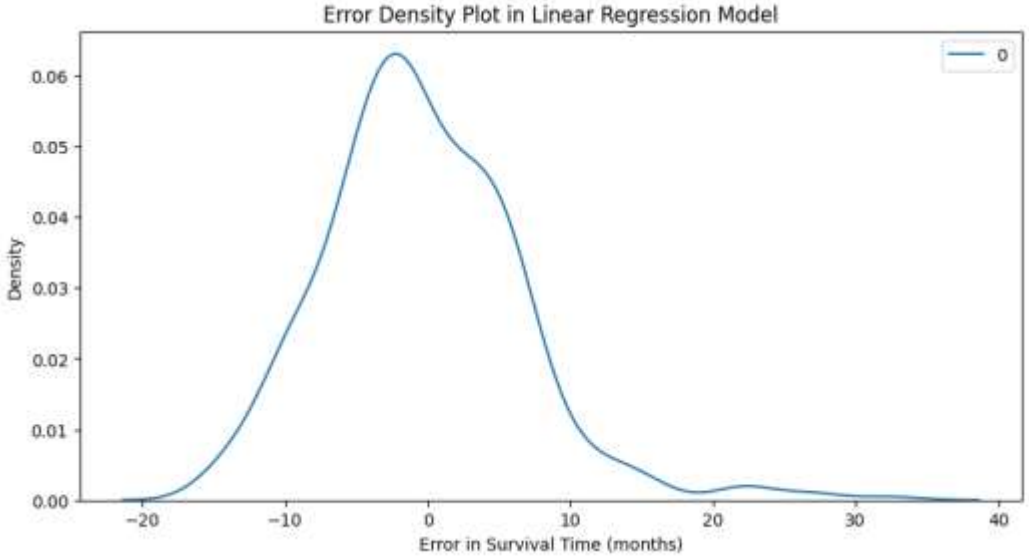
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



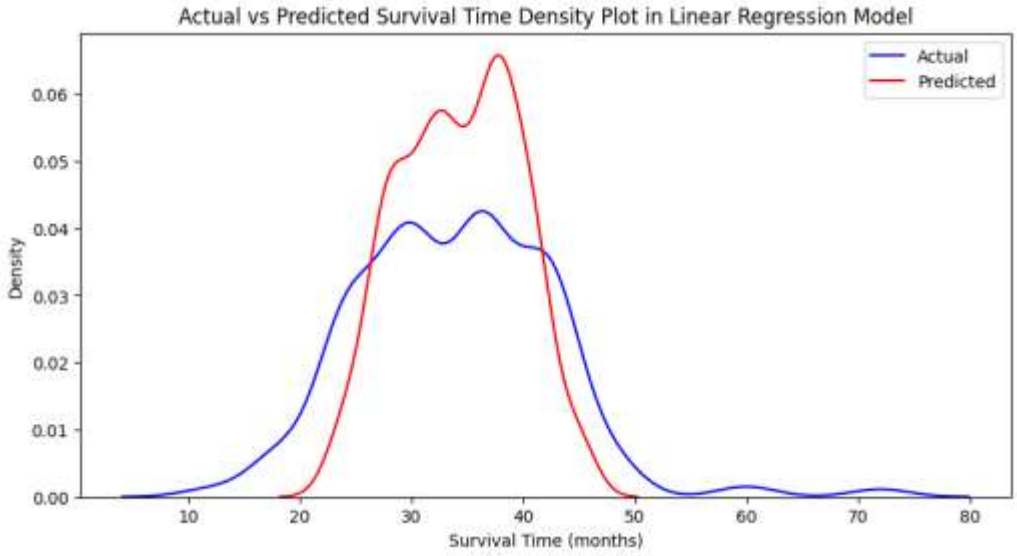
از آنجایی که در رگرسیون خطی در واقع به دنبال به دست آوردن معادله یک خط راست هستیم که به درستی رابطه بین ویژگی‌ها و متغیرهای هدف را نشان دهد، پس در هر معادله برای هر ویژگی یک ضریب توسط مدل آموزش دیده می‌شود که مقادیر ضریب‌ها برای هر ویژگی به شکل زیر است.



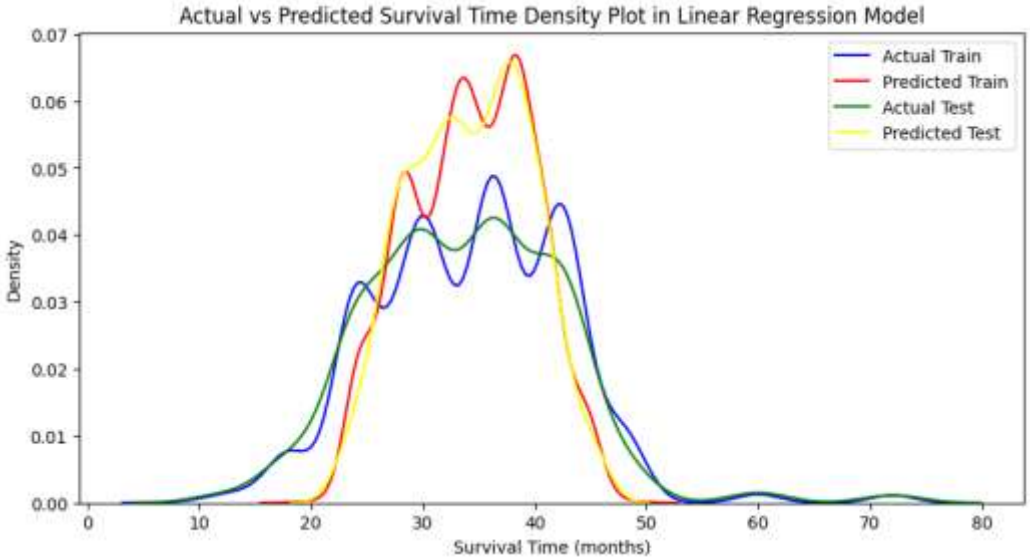
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.



با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، تعداد مقادیر تخمین زده شده برای ماه‌های ۳۵ تا ۴۵ هم در داده‌های آموزش و هم در داده‌های تست بسیار بیشتر از توزیع واقعی این مقادیر است.

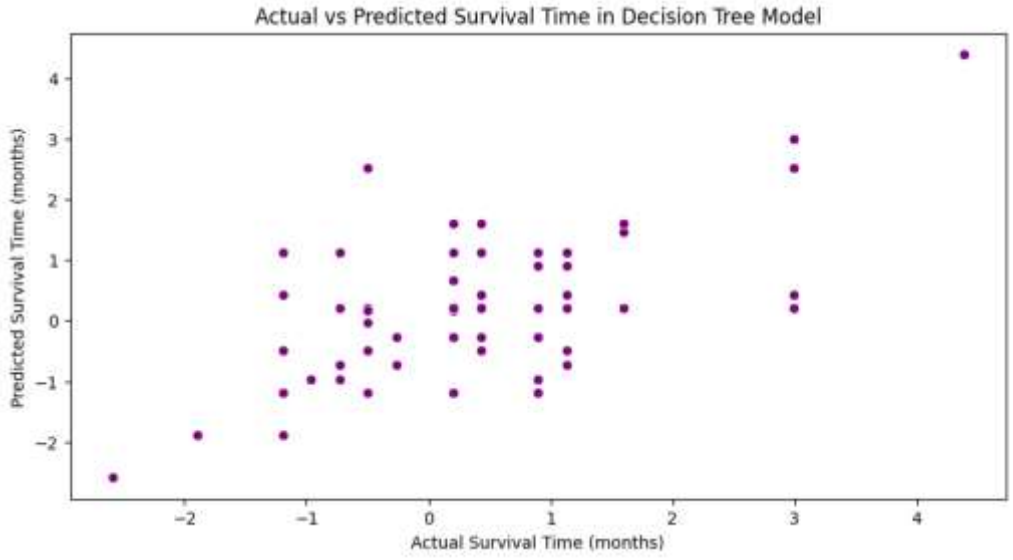


درخت تصمیم

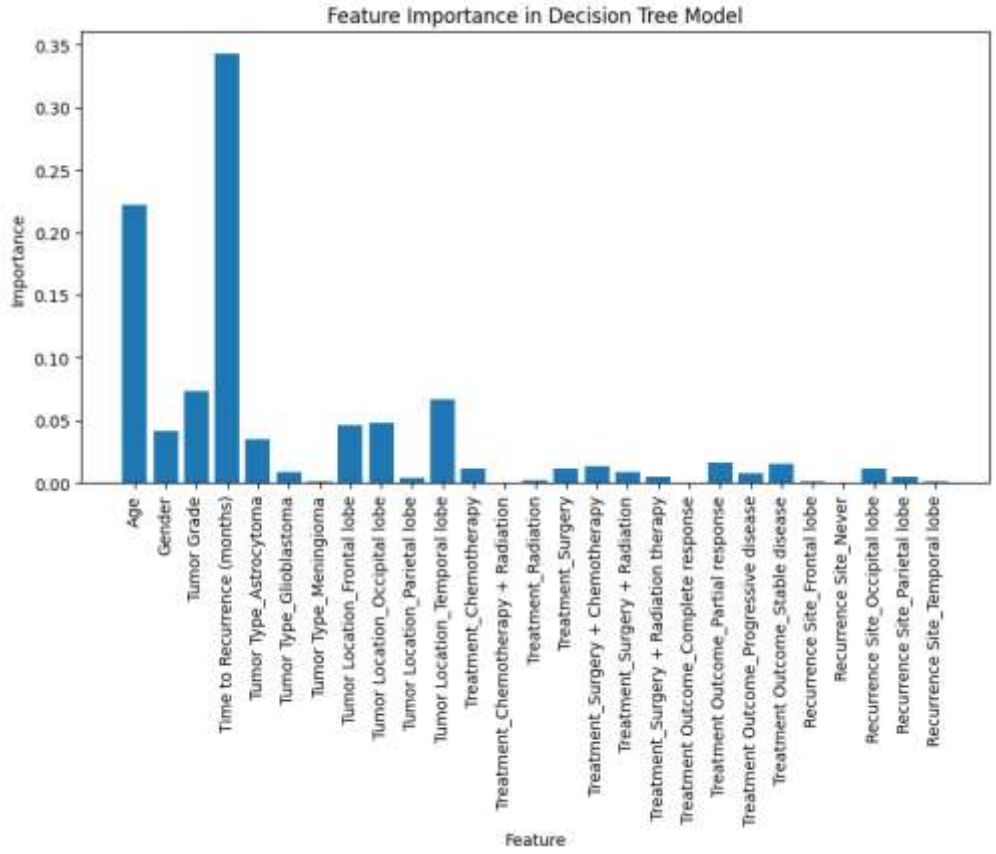
درخت تصمیم رگرسیون یک مدل یادگیری ماشین است که برای پیش‌بینی مقادیر پیوسته استفاده می‌شود. این مدل داده‌ها را به صورت تکراری به بخش‌های کوچکتر تقسیم می‌کند و در هر گره تصمیمی می‌گیرد که کمترین خطای پیش‌بینی را داشته باشد. هر گره داخلی درخت یک ویژگی را برای تقسیم‌بندی انتخاب می‌کند و برگ‌های درخت مقدار پیش‌بینی‌شده نهایی را ارائه می‌دهند. این روش به دلیل سادگی و توانایی در مدل‌سازی روابط غیرخطی محبوب است. با استفاده از ماژول درخت تصمیم رگرسیون در کتابخانه ذکر شده یک مدل درخت تصمیم رگرسیون را با معیار مربعات خطا و عمق دلخواه آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در این روش معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.22767956483444515
Root Mean Squared Error: 0.4771577986729811
R2 Score: 0.7790644051101083
Mean Absolute Error: 0.14652908789492783
```

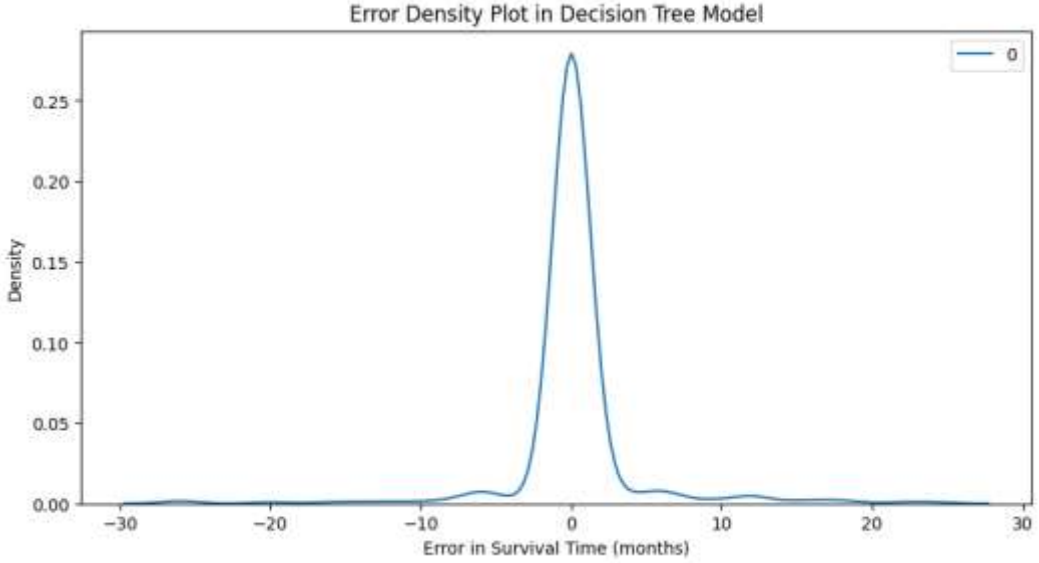
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



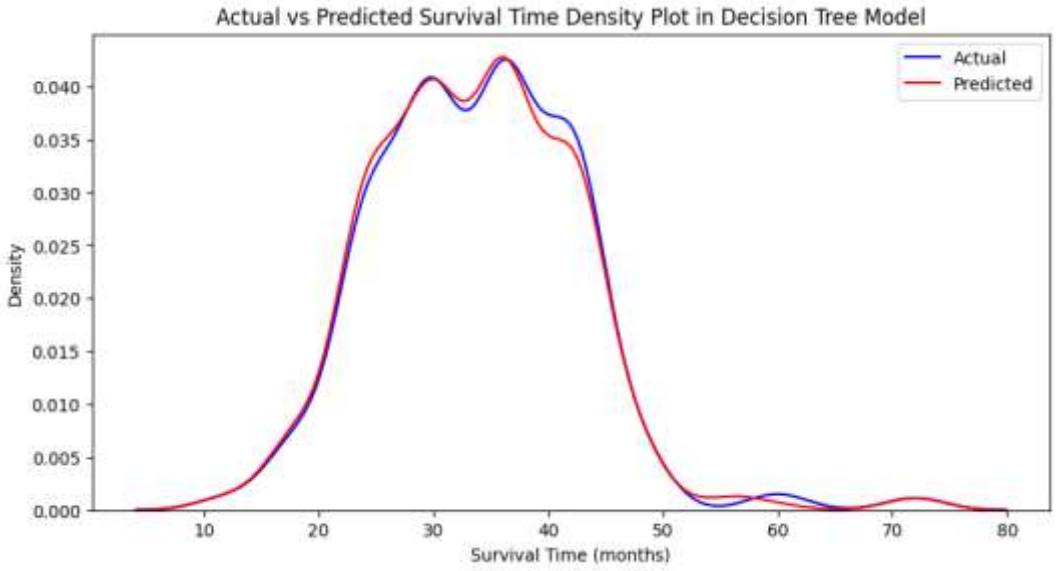
با استفاده از ویژگی feature importances می‌توان میزان اهمیت هر ویژگی را در تصمیم‌گیری به شکل زیر نشان داد.



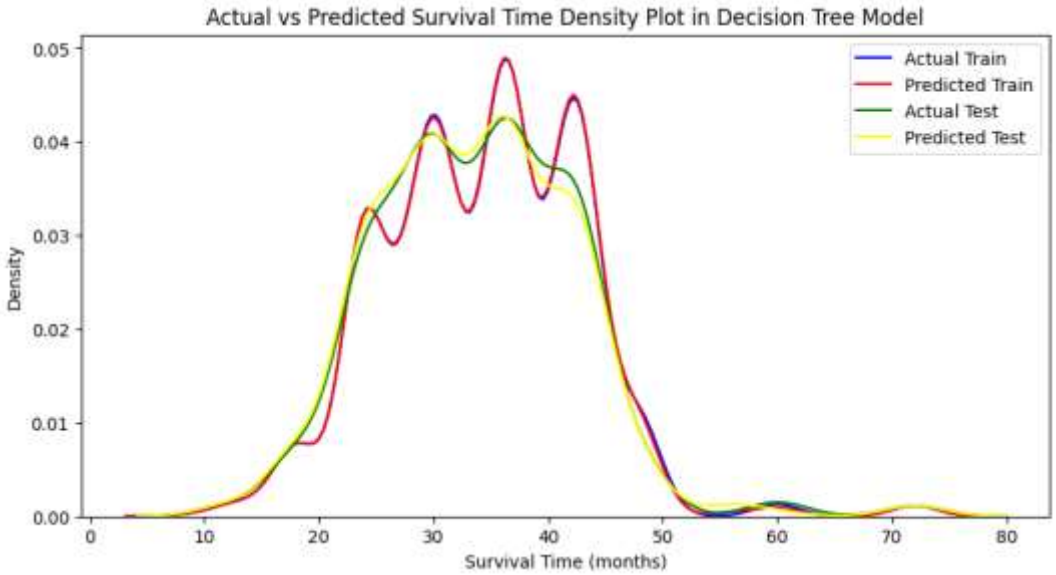
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.



با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، مقادیر واقعی و مقادیر تخمین‌زده شده چگالی نزدیک به هم دارند و می‌توان گفت دقت مدل بسیار خوب است.

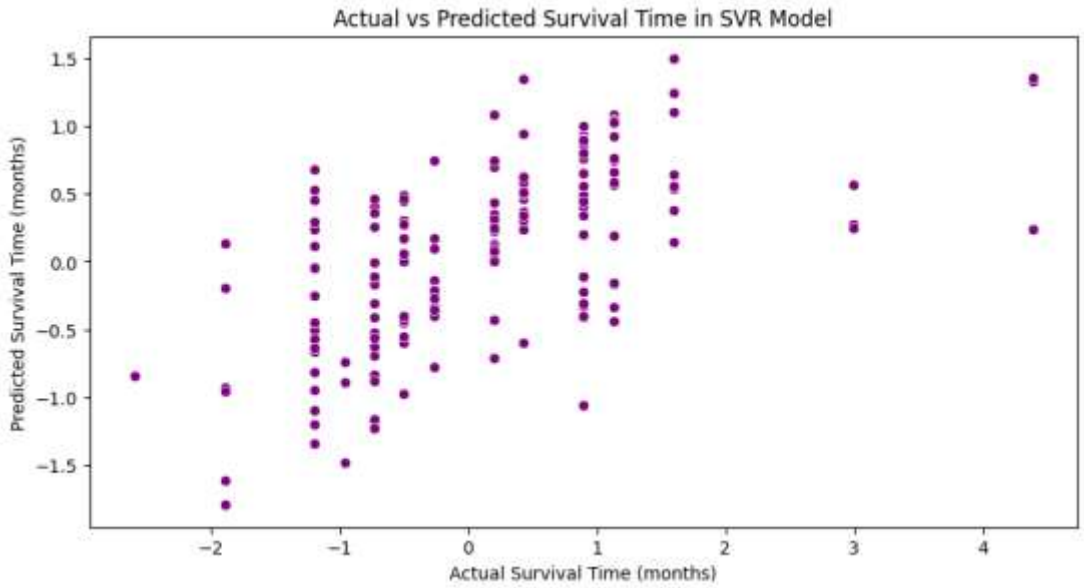


ماشین بردار پشتیبان رگرسیون

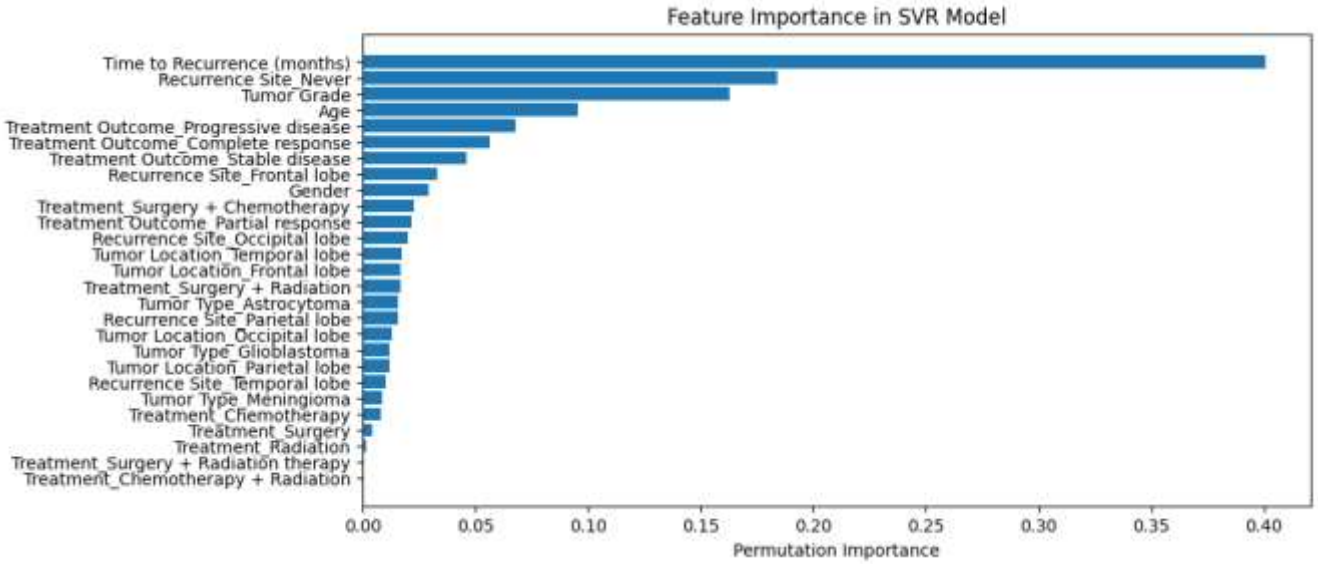
ماشین بردار پشتیبان رگرسیون یکی از روش‌های یادگیری ماشین است که برای پیش‌بینی مقادیر پیوسته استفاده می‌شود. این روش تلاش می‌کند تا یک خط یا یک سطح (در فضای چندبعدی) پیدا کند که به بهترین شکل داده‌ها را براساس یک حاشیه خطا در بر بگیرد. این مدل با تعیین یک حاشیه قابل قبول برای خطاها و یافتن یک تابع با حداکثر شباهت به داده‌ها، سعی می‌کند تا خطای پیش‌بینی را به حداقل برساند. از این روش برای پیش‌بینی‌های دقیق و کارایی بالا در داده‌های پیچیده و با نویز مناسب استفاده می‌شود. با استفاده از ماژول ماشین بردارپ پشتیبان در کتابخانه ذکر شده یک مدل ماشین بردار پشتیبان رگرسیون را با هسته و درجه دلخواه آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در این روش معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.45073811594614166
Root Mean Squared Error: 0.6713703269776984
R2 Score: 0.562612947461836
Mean Absolute Error: 0.3802202288761428
```

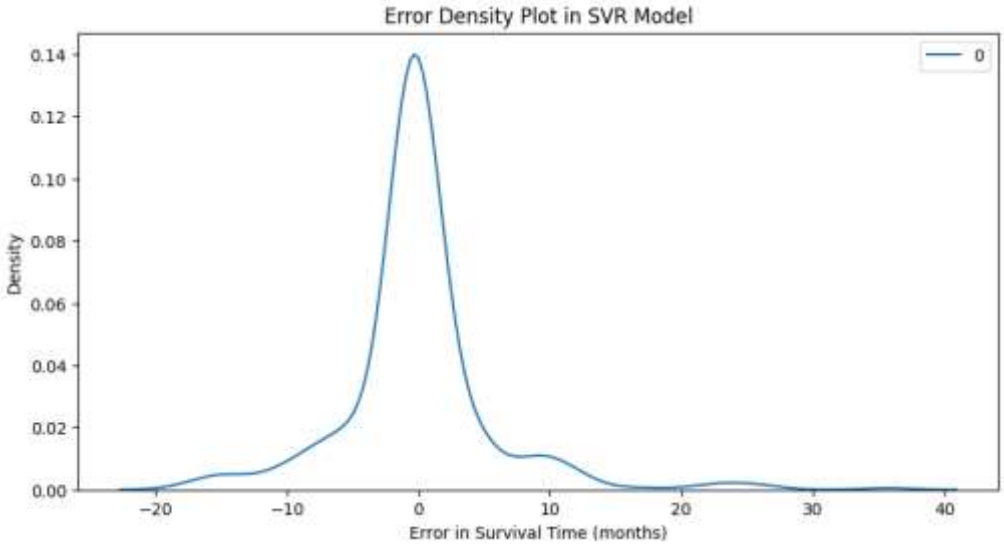
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



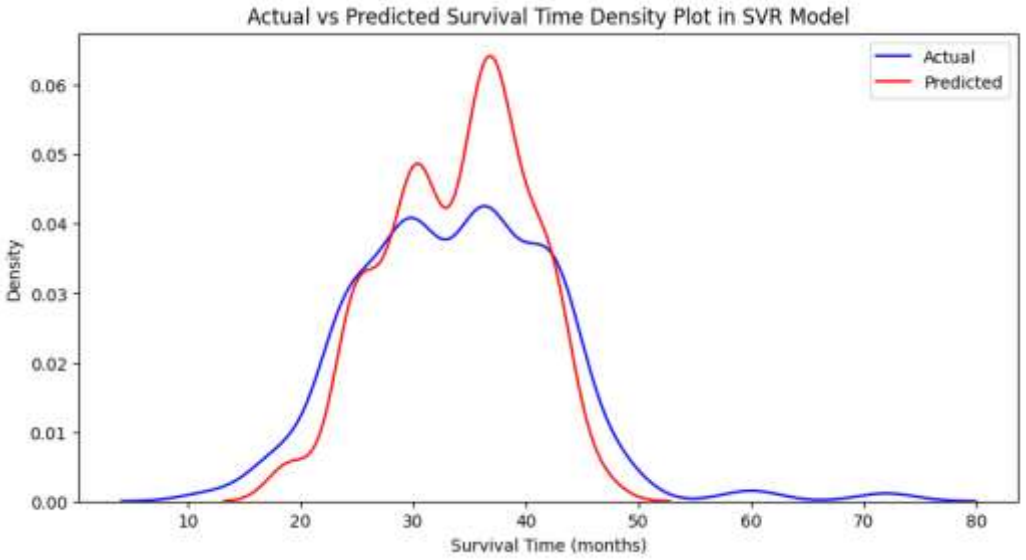
با استفاده از ویژگی feature importances می‌توان میزان اهمیت هر ویژگی را در تصمیم‌گیری به شکل زیر نشان داد.



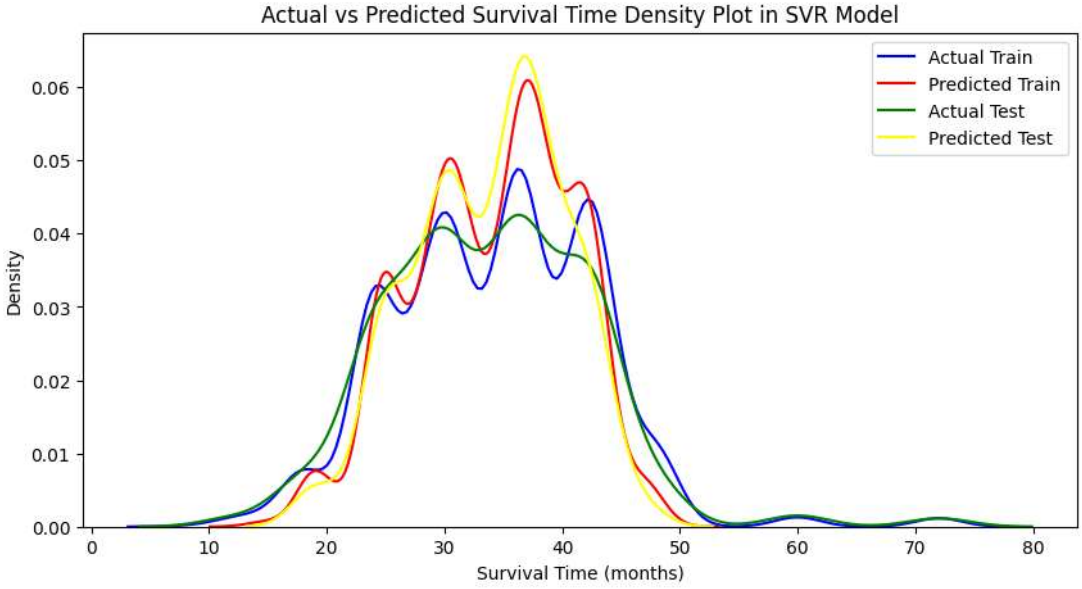
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.



با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، مقادیر واقعی و مقادیر تخمین‌زده شده چگالی نزدیک به هم ندارند و تعداد نمونه داده‌هایی که میزان زنده ماندن برای آن‌ها بین ۳۰ تا ۴۰ ماه پیش‌بینی شده بسیار بیشتر از مقادیر واقعی آن‌ها است. اما نکته قابل توجه این است که الگوی چگالی مقادیر تخمین زده شده تا حدی شبیه مقادیر واقعی است.

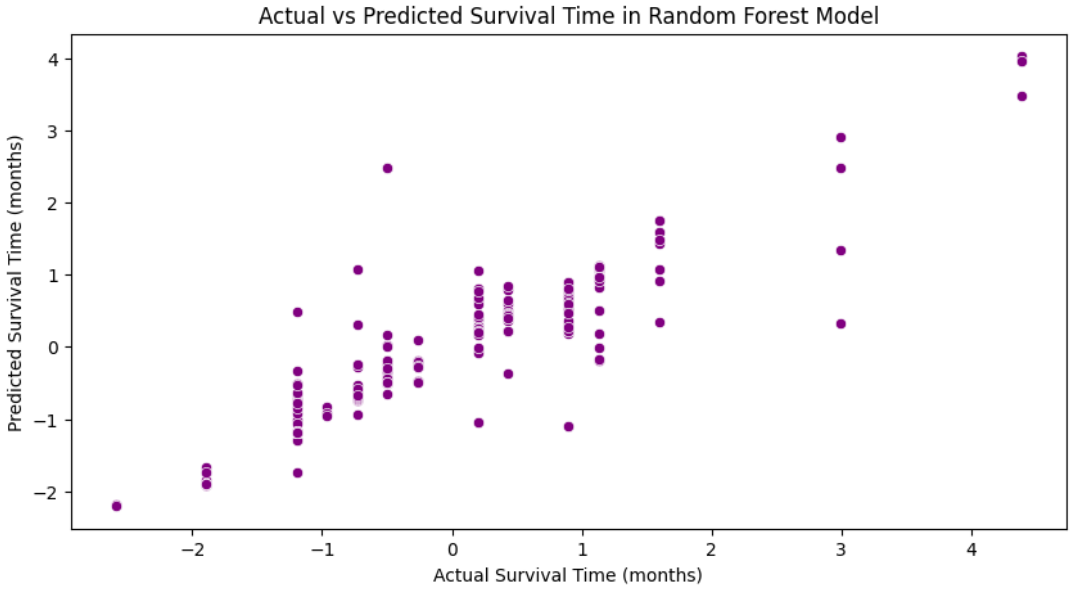


جنگل تصادفی

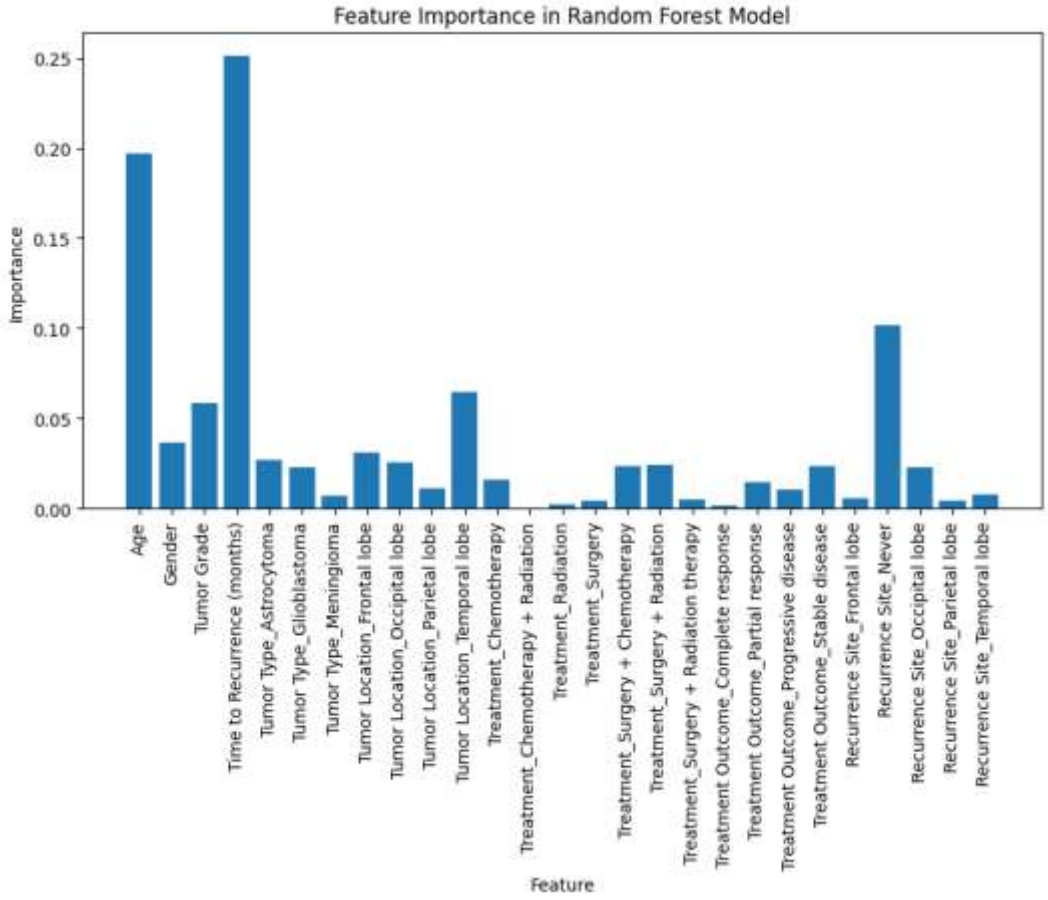
الگوریتم جنگل تصادفی رگرسیون یک روش یادگیری ماشین است که از ترکیب چندین درخت تصمیم برای پیش‌بینی مقادیر پیوسته استفاده می‌کند. این الگوریتم با ساختن تعداد زیادی درخت تصمیم به صورت تصادفی و ترکیب نتایج آنها، دقت پیش‌بینی را افزایش و واریانس مدل را کاهش می‌دهد. هر درخت به‌طور مستقل آموزش داده می‌شود و میانگین پیش‌بینی‌های تمام درختان به عنوان خروجی نهایی ارائه می‌شود. جنگل تصادفی به دلیل مقاومت در برابر بیش‌برازش و توانایی مدیریت داده‌های پیچیده بسیار مورد استفاده قرار می‌گیرد. با استفاده از ماژول جنگل تصادفی رگرسیون در کتابخانه ذکر شده یک مدل جنگل تصادفی رگرسیون را با معیار مربعات خطا و تعداد درخت دلخواه آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در این روش معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.17701218952008943
Root Mean Squared Error: 0.4207281658269261
R2 Score: 0.8282309902391969
Mean Absolute Error: 0.17096030359586567
```

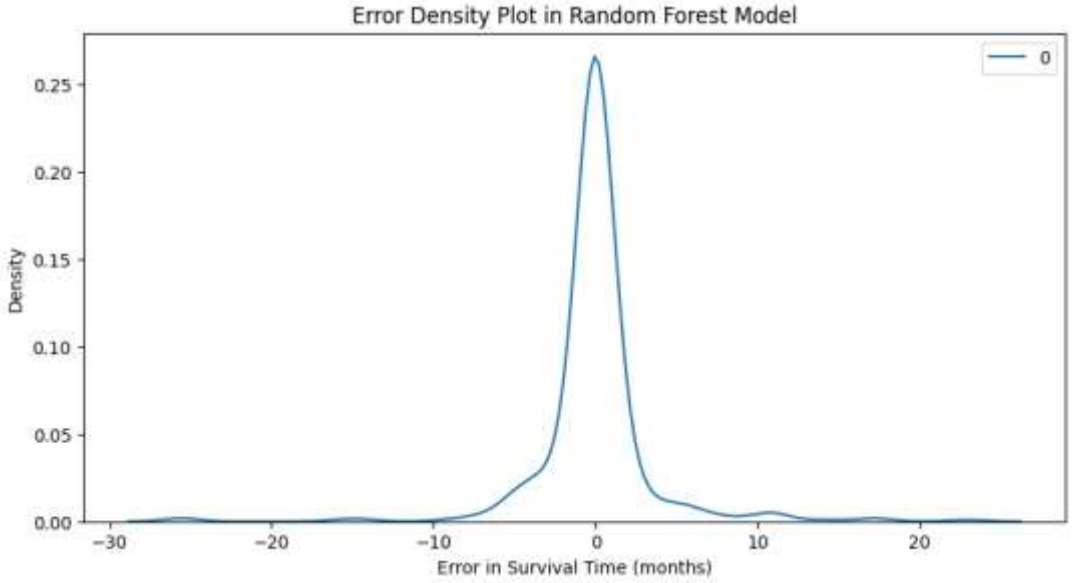
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



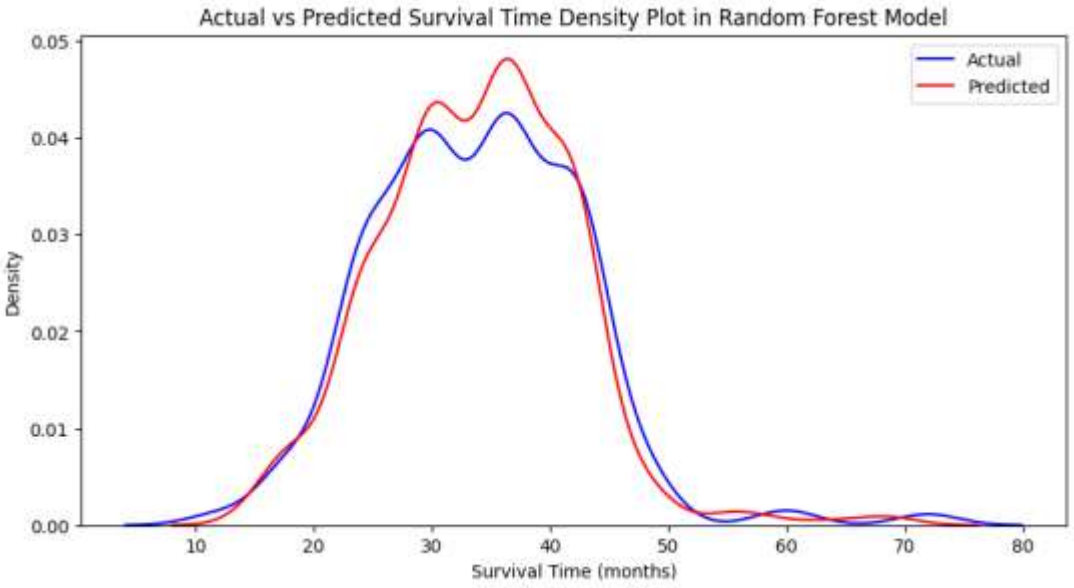
با استفاده از ویژگی feature importances می‌توان میزان اهمیت هر ویژگی را در تصمیم‌گیری به شکل زیر نشان داد.



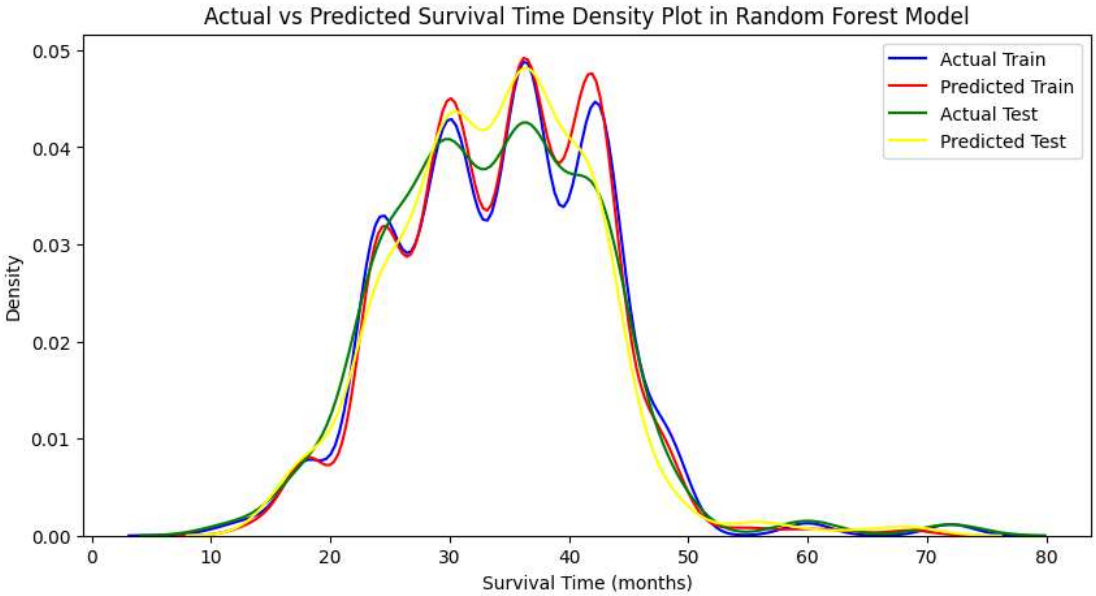
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.



با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، مقادیر واقعی و مقادیر تخمین‌زده شده با حفظ تقریبی الگو مقداری از هم فاصل دارند.

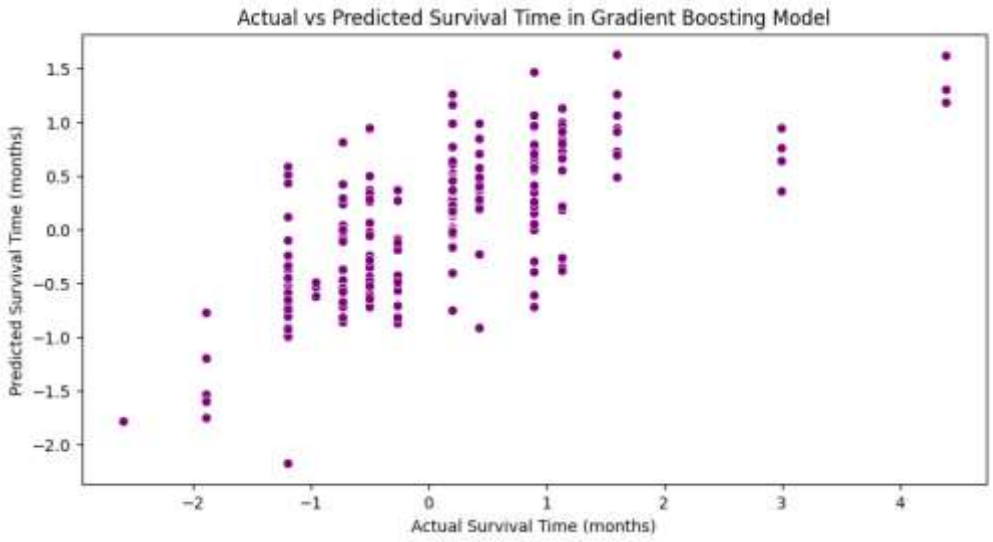


تقویت گرادیان

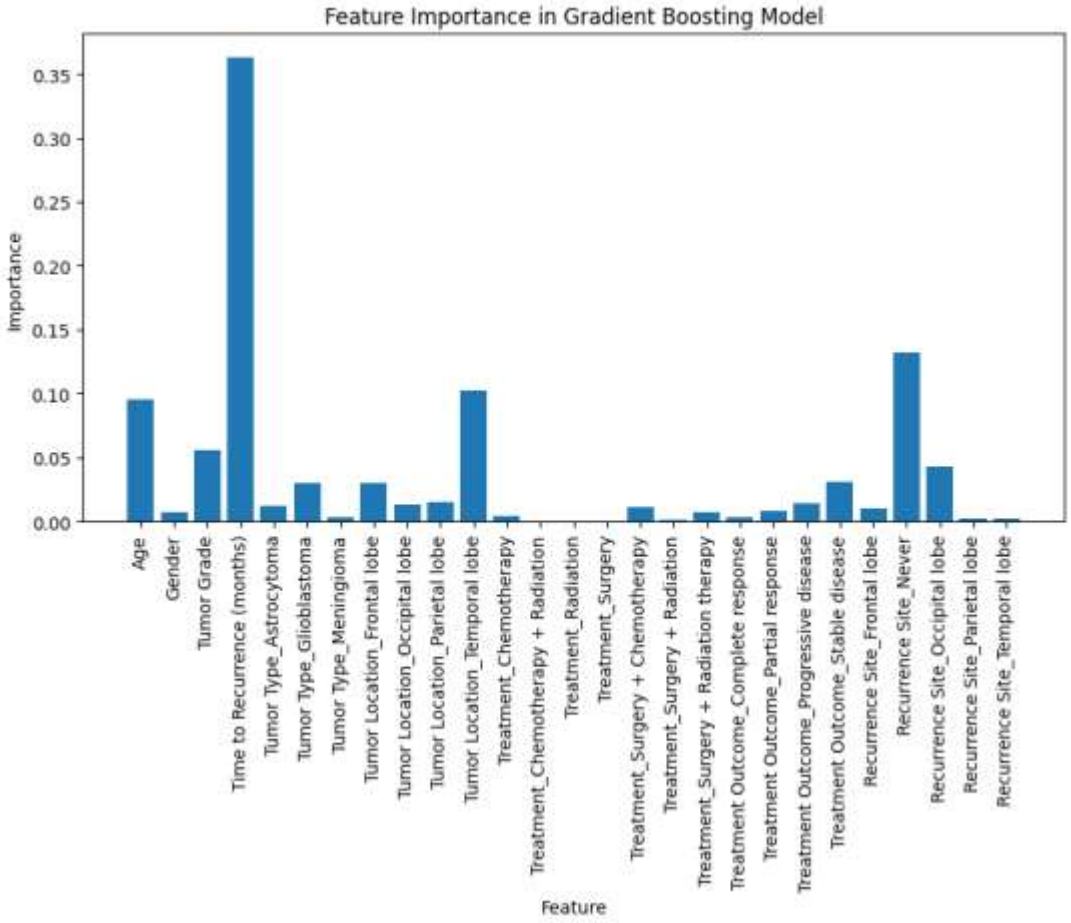
الگوریتم تقویت گرادیان رگرسیون یک تکنیک یادگیری ماشین است که برای بهبود دقت مدل‌های رگرسیون استفاده می‌شود. این الگوریتم با ترکیب تعداد زیادی مدل ضعیف (اغلب درخت‌های تصمیم ساده) به صورت متوالی، یک مدل قوی‌تر ایجاد می‌کند. در هر مرحله، مدل جدیدی آموزش داده می‌شود تا خطاهای مدل قبلی را اصلاح کند. با استفاده از ماژول تقویت گرادیان رگرسیون در کتابخانه ذکر شده یک مدل تقویت گرادیان رگرسیون را با معیار مربعات خطا و تعداد درخت دلخواه آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در این روش معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.39047424908696504
Root Mean Squared Error: 0.6248793876316974
R2 Score: 0.621091771789258
Mean Absolute Error: 0.420562680761752
```

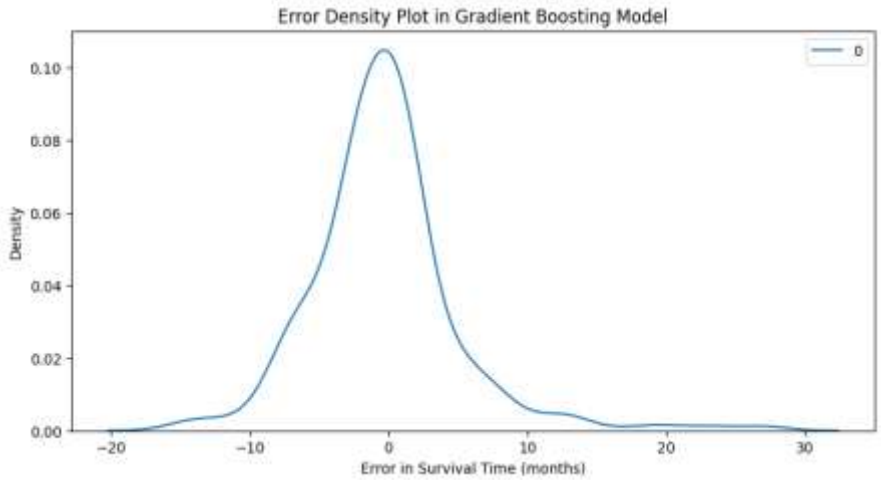
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



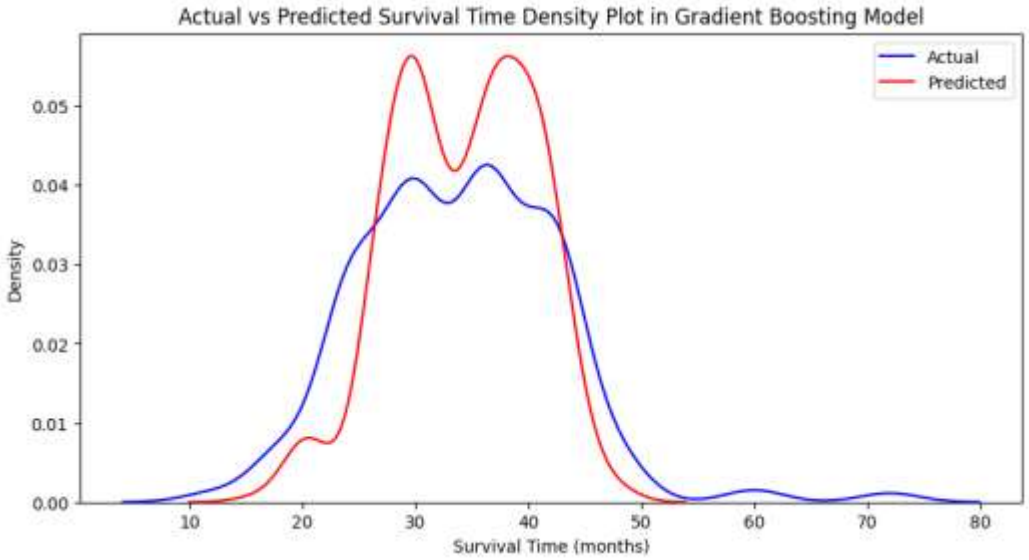
با استفاده از ویژگی feature importances می‌توان میزان اهمیت هر ویژگی را در تصمیم‌گیری به شکل زیر نشان داد.



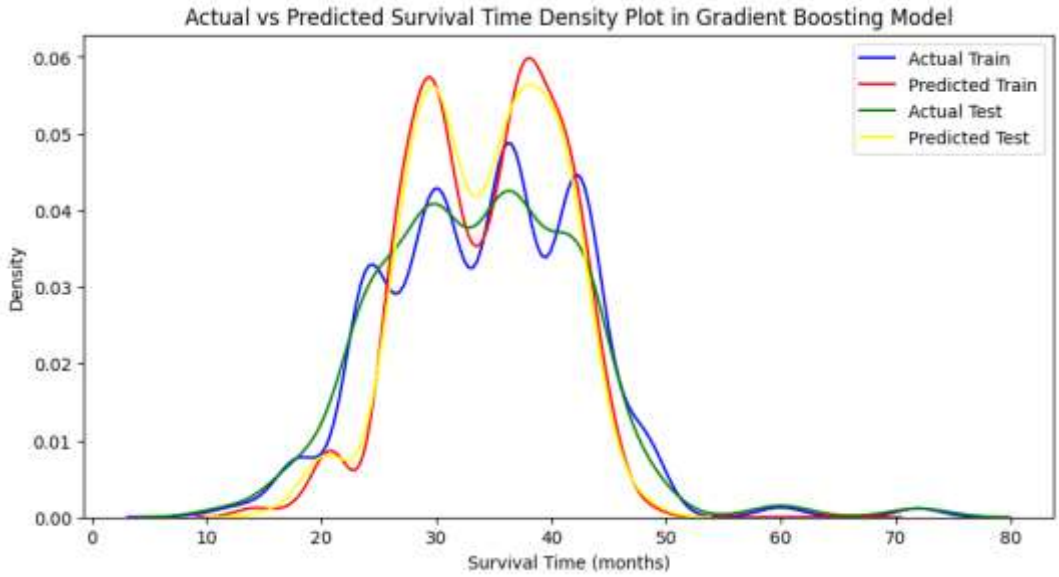
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.



با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، مقادیر واقعی و مقادیر تخمین‌زده شده بدون حفظ الگو مقدار زیادی از هم فاصل دارند.

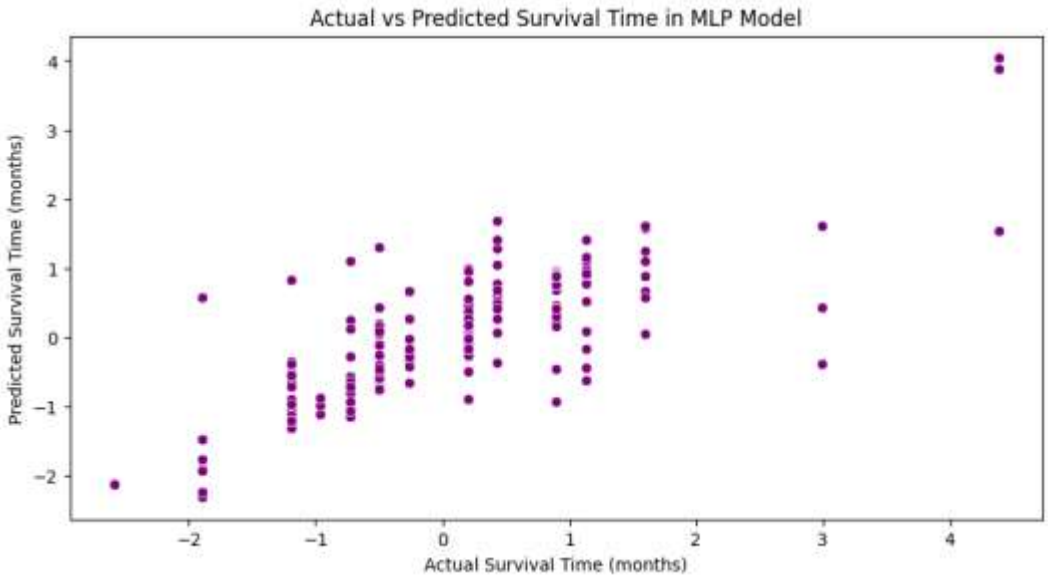


پرسپترون چند لایه

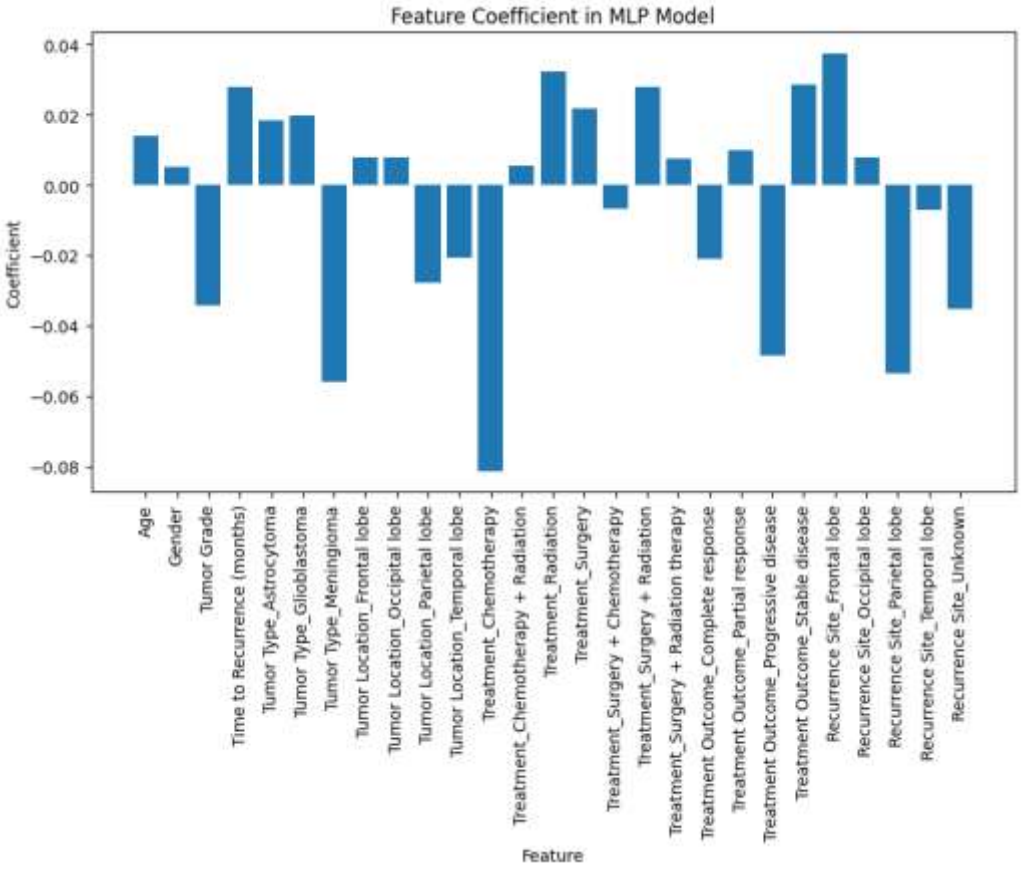
الگوریتم پرسپترون چند لایه یک نوع شبکه عصبی مصنوعی است که برای مسائل رگرسیون و دسته‌بندی استفاده می‌شود. این الگوریتم شامل چندین لایه از نورون‌ها است: یک لایه ورودی، یک یا چند لایه پنهان، و یک لایه خروجی. هر نورون به نورون‌های لایه قبلی و بعدی متصل است و از توابع فعال‌ساز (ReLU و تابع همانی برای رگرسیون) برای یادگیری روابط پیچیده بین ورودی‌ها و خروجی‌ها استفاده می‌کند. این روش با استفاده از پس‌انتشار خطا آموزش داده می‌شود و توانایی بالایی در مدل‌سازی داده‌های غیرخطی و پیچیده دارد. با استفاده از ماژول پرسپترون چند لایه رگرسیون در کتابخانه ذکر شده یک مدل پرسپترون چند لایه را با تابع فعال‌ساز و تعداد نورون‌های پنهان دلخواه آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در این روش معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.26736923499485543
Root Mean Squared Error: 0.5170775908844392
R2 Score: 0.7405503606272403
Mean Absolute Error: 0.26667148853475153
```

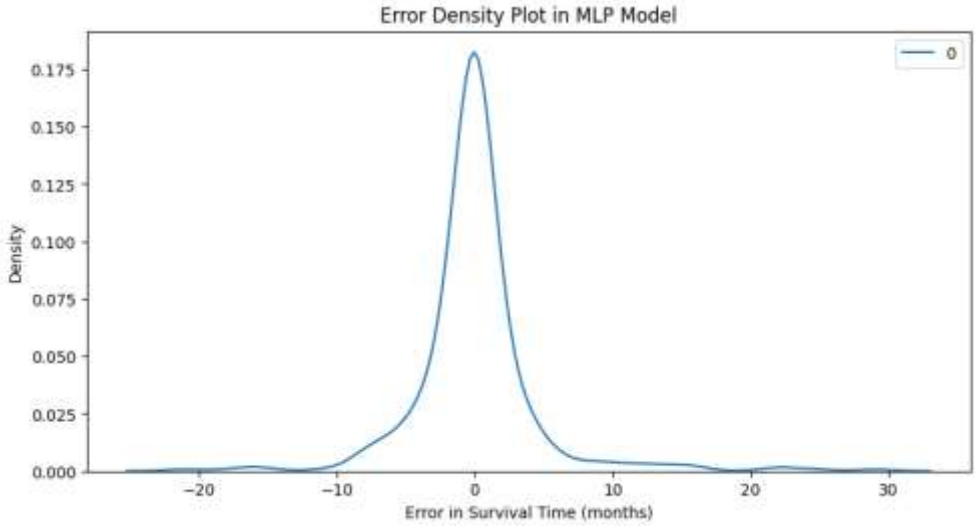
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



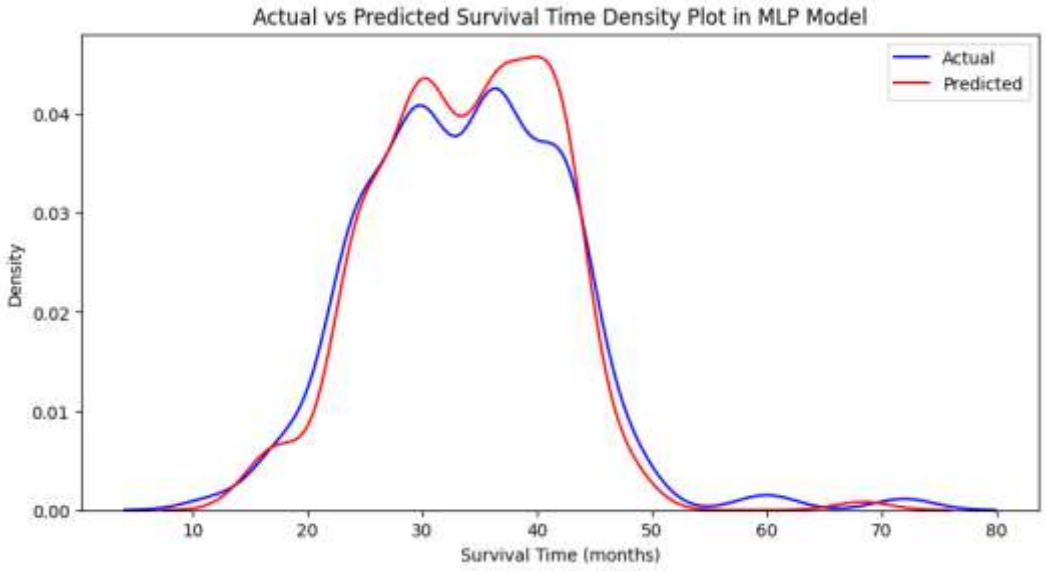
در نمودار زیر می‌توان ضریب مربوط به هر یک از ویژگی‌ها را برای انجام پیش‌بینی توسط مدل مشاهده کرد.



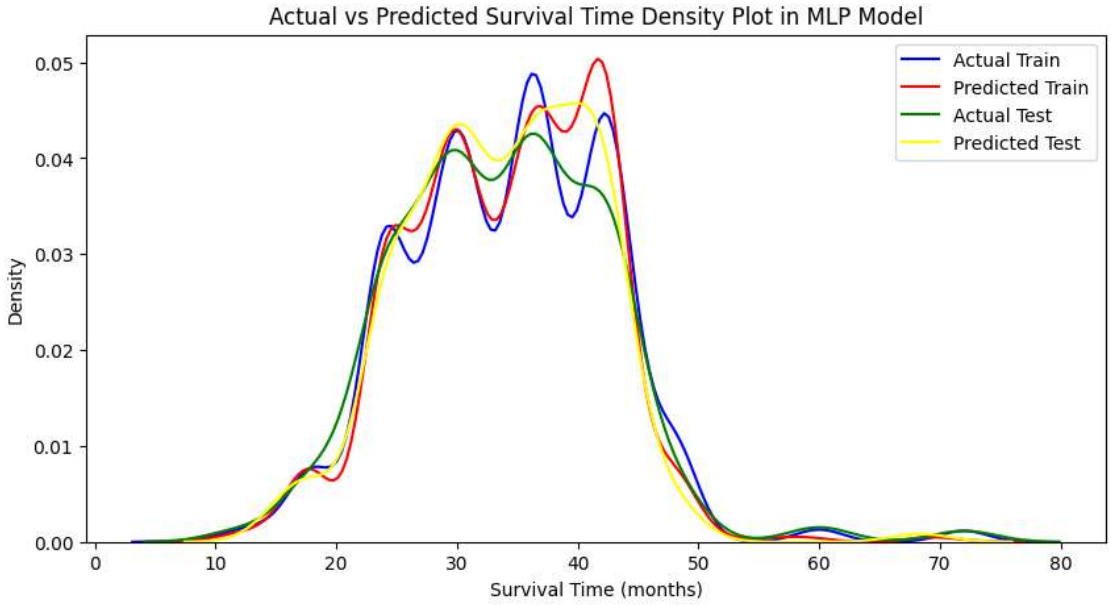
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.



با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، مقادیر واقعی و مقادیر تخمین‌زده شده بدون حفظ الگو مقدار کمی از هم فاصله دارند.

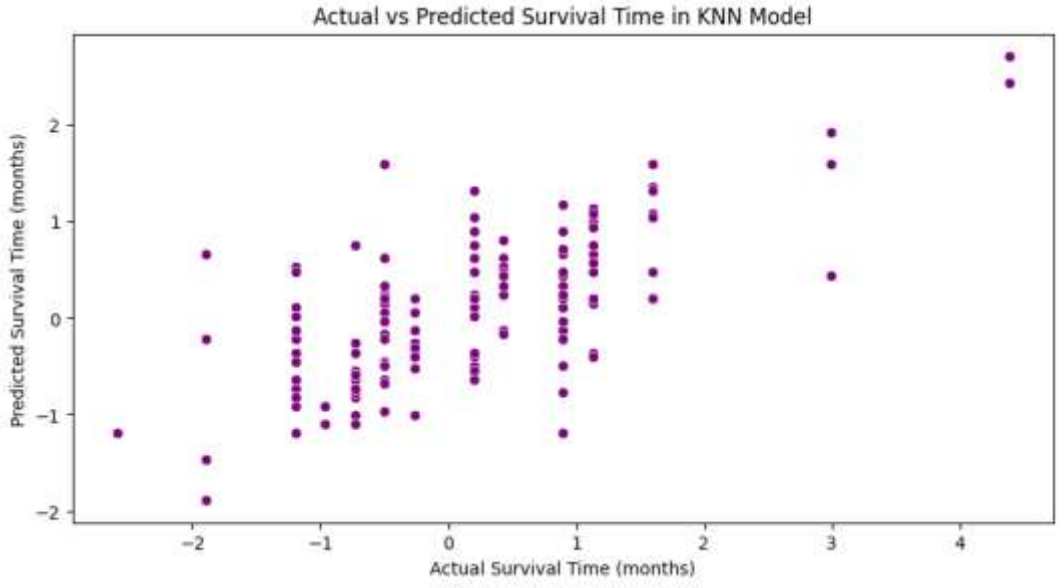


K نزدیک‌ترین همسایه

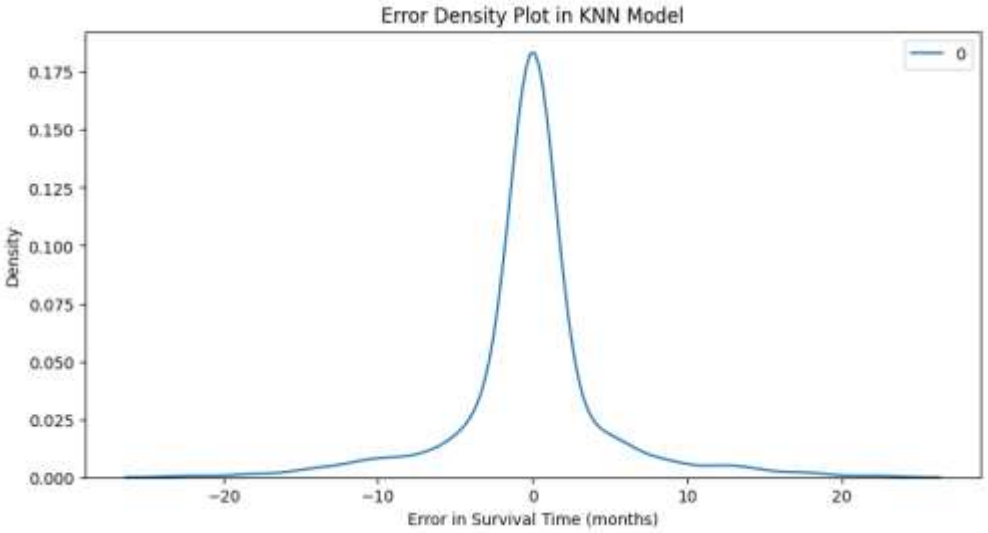
الگوریتم k نزدیک‌ترین همسایه در رگرسیون به این صورت عمل می‌کند که بر اساس ویژگی‌های ورودی (مشخصه‌ها) مانند فاصله اقلیدسی یا فاصله منهتن، k نمونه از داده‌های آموزشی که به نزدیک‌ترین ویژگی‌های ورودی نزدیکترین همسایه‌ها هستند را انتخاب می‌کند. سپس برای پیش‌بینی خروجی مورد نظر، میانگین مقادیر خروجی متناظر با این k نمونه محاسبه می‌شود. این الگوریتم به عنوان یک روش ساده و موثر برای پیش‌بینی مقادیر پیوسته استفاده می‌شود و معمولاً در مواردی که توزیع داده‌ها یا رابطه بین ورودی و خروجی پیچیده‌تر نیست، عملکرد خوبی دارد. با استفاده از ماژول k نزدیک‌ترین همسایه رگرسیون در کتابخانه ذکر شده یک مدل k نزدیک‌ترین همسایه رگرسیون را با وزن‌ها و تعداد نزدیک‌ترین همسایه دلخواه آموزش داده و معیارهای میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، امتیاز R2 و میانگین قدرمطلق خطا (MAE) را بر روی داده‌های تست ارزیابی می‌کنیم. در این روش معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Mean Squared Error: 0.3069992611313201
Root Mean Squared Error: 0.5540751403296489
R2 Score: 0.7020941934858091
Mean Absolute Error: 0.27823650645342
```

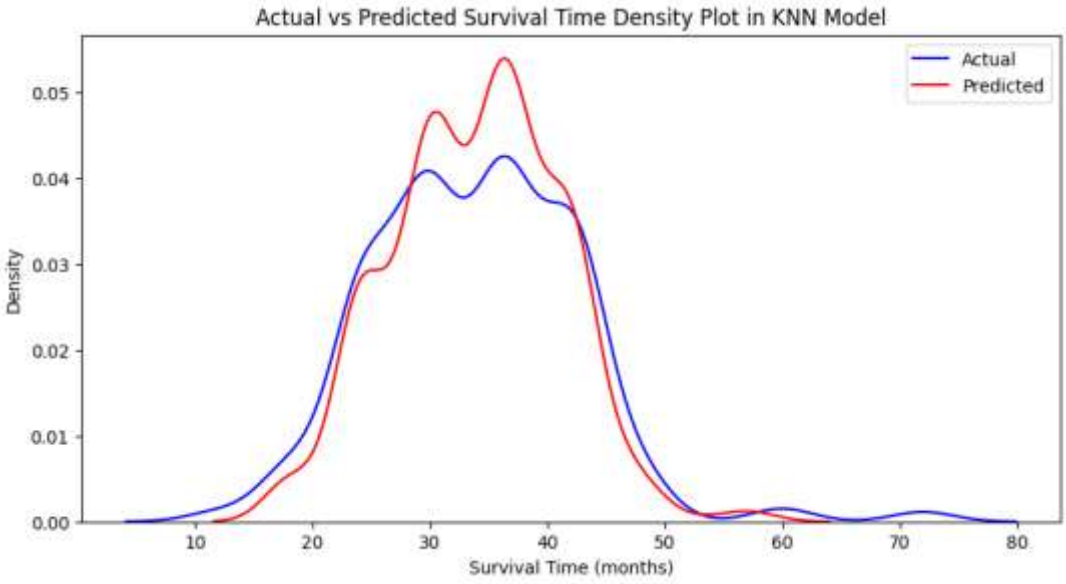
حال، می‌توان نمودار پراکندگی مقادیر واقعی و مقادیر پیش‌بینی شده توسط مدل برای داده‌های تست را به صورت زیر رسم کرد. در صورتی که پیش‌بینی مدل دقت بالایی داشته باشد مقدار هر نقطه در محور عمودی و افقی با یکدیگر برابرخواهد بود.



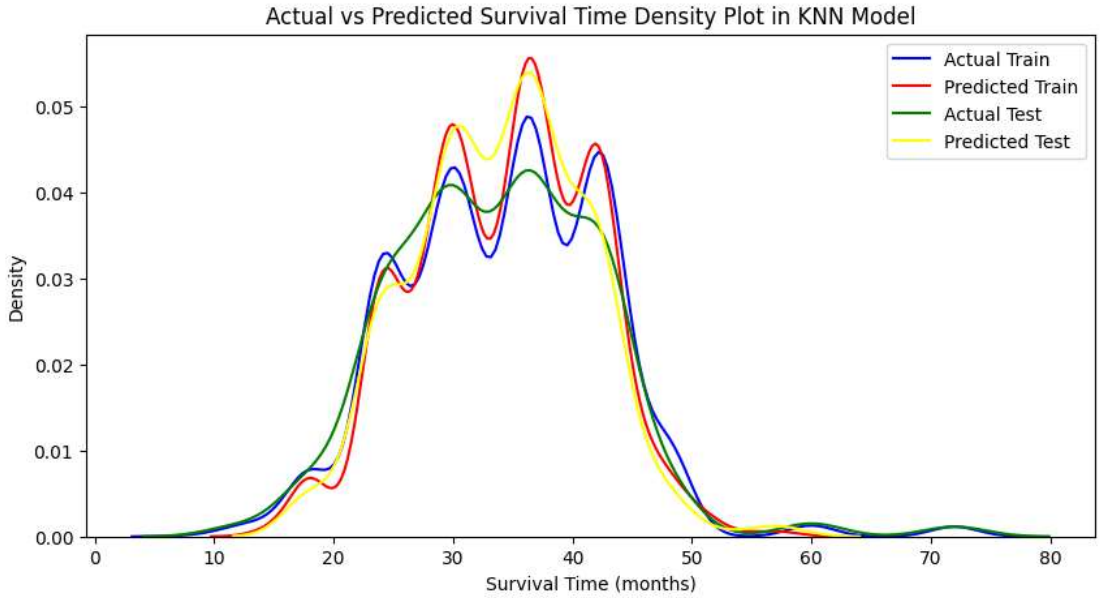
اکنون میزان خطای مدل را با توجه به مقدار واقعی و مقدار پیش‌بینی شده مدل بررسی می‌کنیم. هر چه میزان خطا به صفر نزدیک‌تر باشد، مدل عملکرد بهتری داشته است.



در این مرحله، با توجه به نمودار زیر می‌توان چگالی تعداد ماه‌های واقعی و تخمین زده شده زنده ماندن افراد را در نمونه داده تست مشاهده کرد. هر چه خط مقدار پیش‌بینی شده توسط مدل به نمودار مقادیر واقعی نزدیک‌تر باشد، عملکرد مدل بهتر خواهد بود.

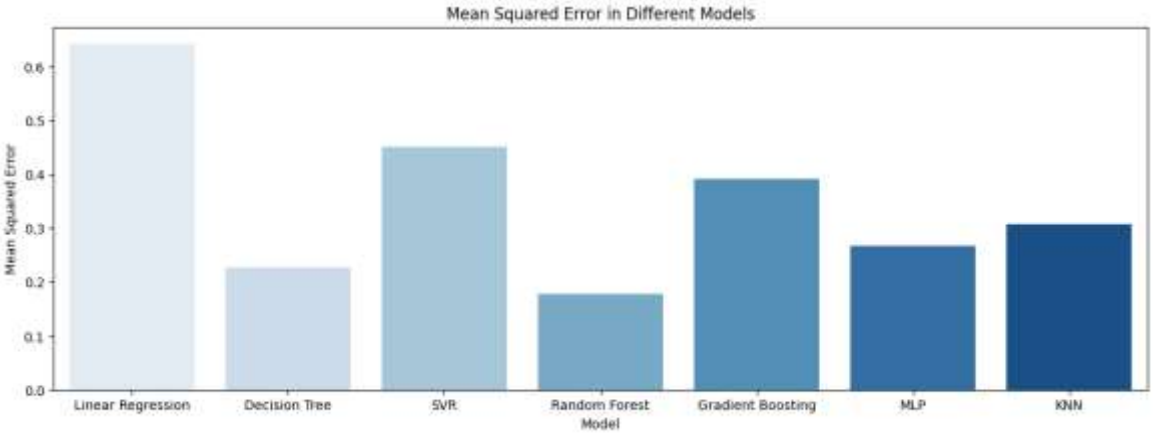


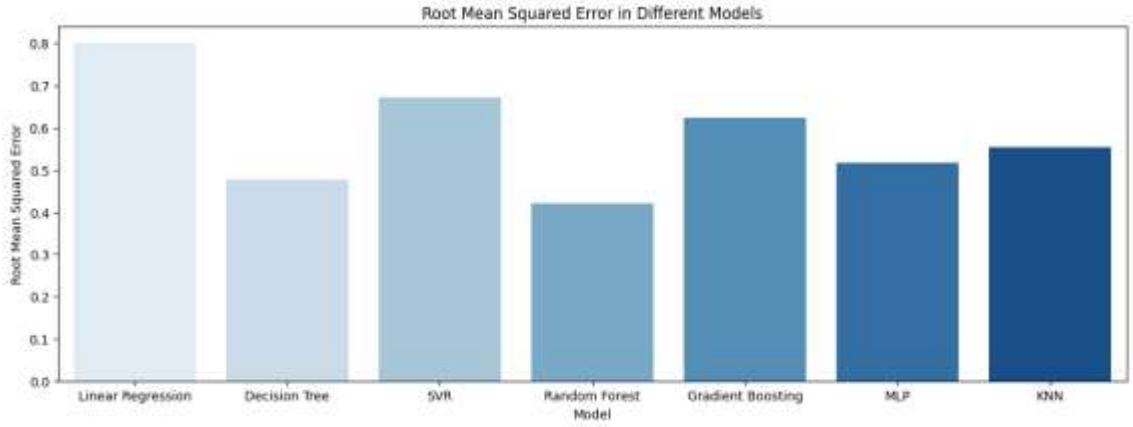
با تعمیم نمودار بالا برای داده‌های آموزش نیز می‌توان چنین نموداری را رسم کرد. همانطور که مشاهده می‌شود، مقادیر واقعی و مقادیر تخمین‌زده شده با حفظ الگو مقدار زیادی از یکدیگر فاصله دارند.

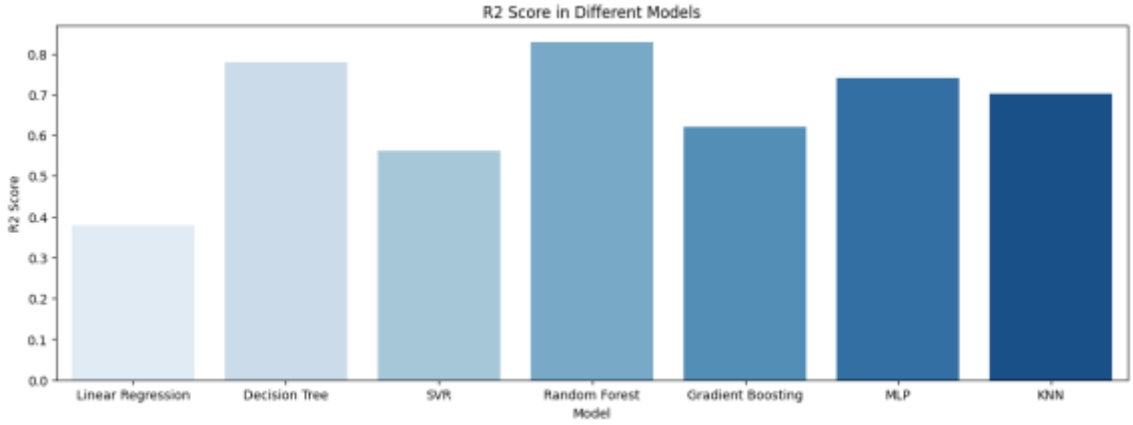


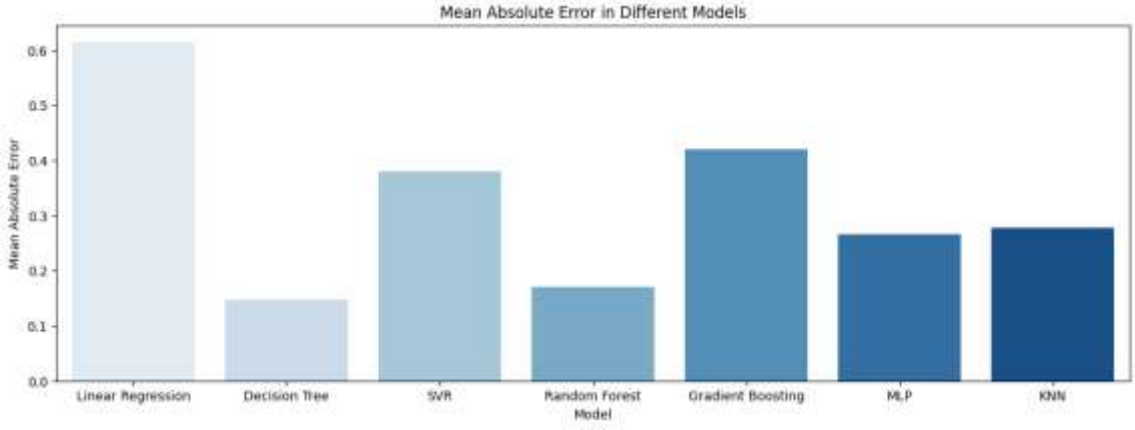
مقایسه مدل‌های یادگیری ماشین

در این بخش با استفاده از معیارهای مشخص شده و همچنین نمودارهای رسم شده به بررسی عملکرد روش‌های مختلف می‌پردازیم. در نمودارهای زیر مشاهده می‌شود رگرسیون خطی با توجه به چهار معیار خطای بیشتر و عملکرد ضعیف‌تری داشته است. این در حالی است که الگوریتم جنگل تصادفی کمترین میزان خطا و بیشترین میزان امتیاز R2 را داشته است و در رتبه دوم الگوریتم درخت تصمیم قرار دارد. لازم به ذکر است فضای نمونه ابرپارامترهای مدل بسیار بزرگ بوده و با تعیین ابر پارامترهای بهینه، می‌توان به مدل‌هایی با عملکرد بهتر دست یافت.

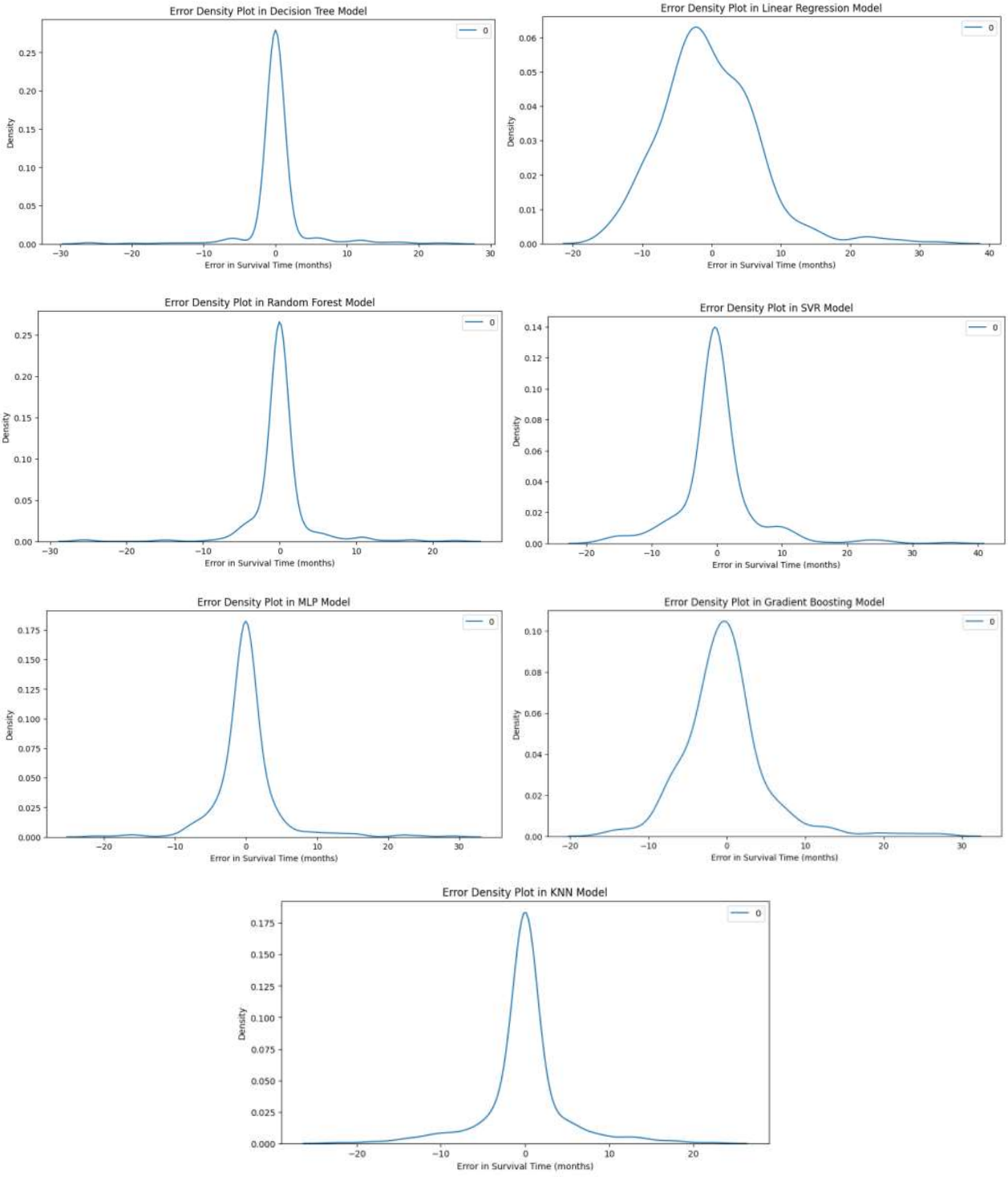



Root Mean Squared Error in Different Models
Root Mean Squared Error
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Linear Regression
Decision Tree
SVR
Random Forest
Gradient Boosting
MLP
KNN
Model

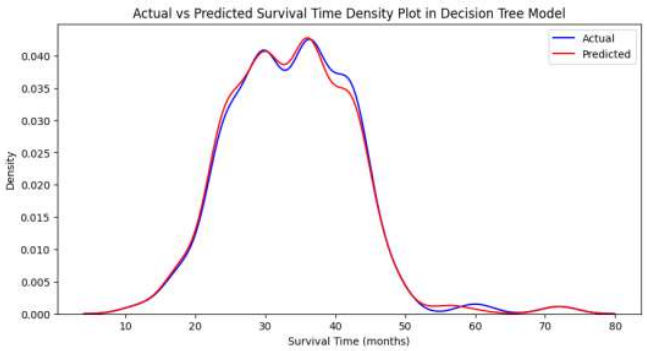

R2 Score in Different Models
R2 Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Linear Regression
Decision Tree
SVR
Random Forest
Gradient Boosting
MLP
KNN
Model

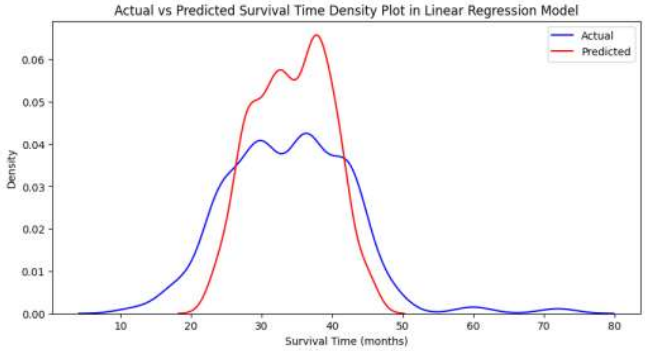

Mean Absolute Error in Different Models
Mean Absolute Error
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Linear Regression
Decision Tree
SVR
Random Forest
Gradient Boosting
MLP
KNN
Model

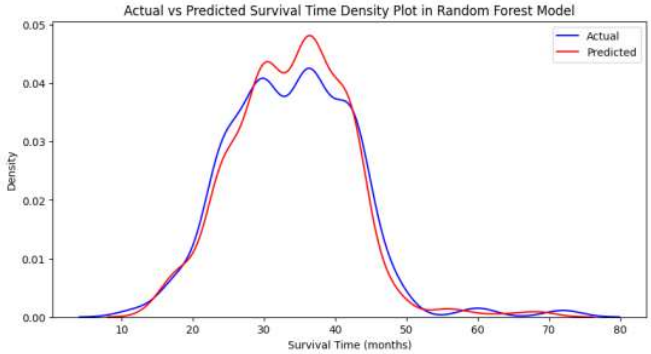
نتیجه‌گیری بالا نیز با کنار هم قرار دادن نمودارهای چگالی خطای هر مدل قابل تایید است.

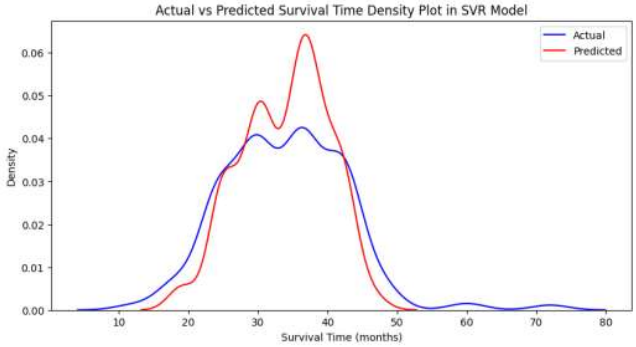


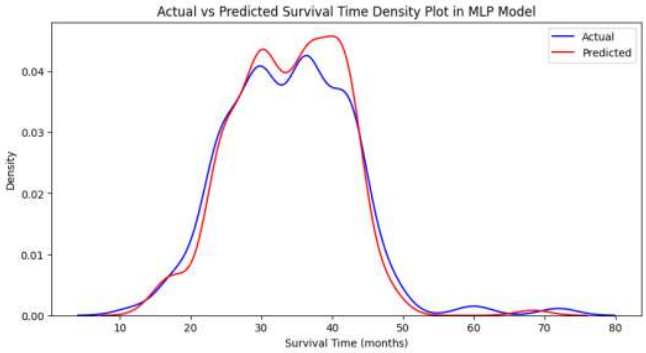
همچنین با استفاده از نمودارهای چگالی مقدار واقعی متغیر هدف با مقدار تخمین زده شده، می‌توان به همان نتیجه رسید که الگوریتم‌های جنگل تصادفی و درخت تصمیم رگرسیون بهترین عملکردها را داشته‌اند.

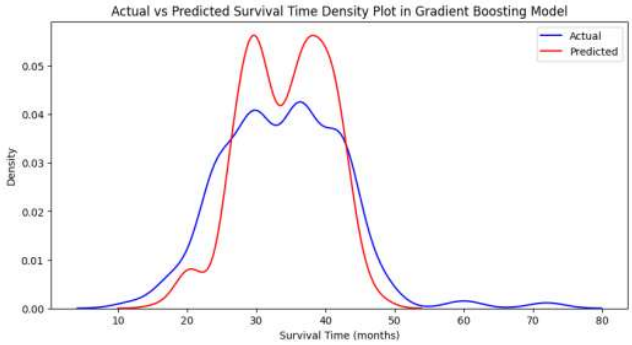


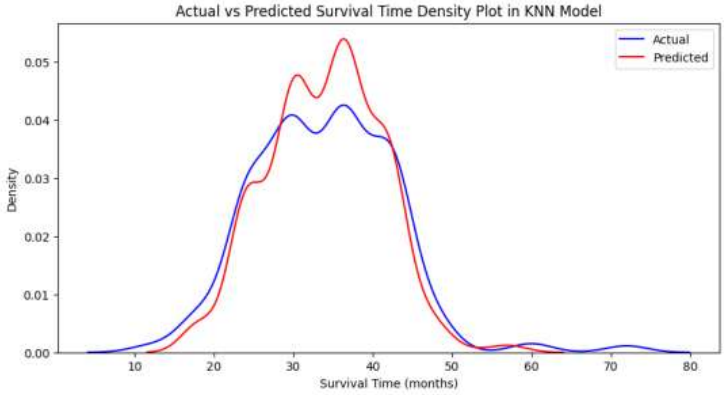












مسئله دسته‌بندی: زنده ماندن یا نماندن بیماران دارای کوید ۱۹

دسته‌بندی یکی از روش‌های یادگیری ماشین است که برای تخصیص نمونه‌های داده به یکی از چندین دسته یا گروه از پیش تعریف‌شده استفاده می‌شود. این تکنیک در مواردی به کار می‌رود که خروجی به صورت دسته‌بندی شده باشد، مانند تشخیص بیماری (بیمار/سالم)، شناسایی اسپم در ایمیل (اسپم/غیر اسپم)، و پیش‌بینی رفتار مشتری (خرید خواهد کرد/خرید نخواهد کرد). بیماری کرونا (کووید-۱۹) یک بیماری عفونی است که توسط یک ویروس کرونا که به تازگی کشف شده بود، ایجاد می‌شود. اکثر افرادی که به ویروس کووید-۱۹ مبتلا می‌شوند، دچار بیماری تنفسی خفیف تا متوسط شده و بدون نیاز به درمان خاصی بهبود می‌یابند. افراد مسن و کسانی که دارای مشکلات زمینه‌ای پزشکی مانند بیماری‌های قلبی عروقی، دیابت، بیماری‌های مزمن تنفسی و سرطان هستند، بیشتر در معرض ابتلا به بیماری شدید قرار دارند. در طول همه‌گیری بیماری، یکی از مشکلات اصلی که ارائه دهندگان خدمات بهداشتی با آن روبرو بوده‌اند، کمبود منابع پزشکی و فقدان برنامه‌ای مناسب برای توزیع کارآمد آن‌ها بوده است. در این دوران سخت، توانایی پیش‌بینی اینکه یک فرد در زمان آزمایش مثبت یا حتی قبل از آن به چه نوع منابعی نیاز دارد، کمک بزرگی به مقامات خواهد کرد، زیرا آن‌ها قادر خواهند بود تا منابع ضروری را برای نجات جان آن بیمار تهیه و سازماندهی کنند. مسئله دسته‌بندی در این پروژه به استفاده از مجموعه داده‌های کووید-۱۹ برای پیش‌بینی وضعیت زنده ماندن یا نماندن بیماران مبتلا به این بیماری می‌پردازد. با تحلیل ویژگی‌های مختلف بیماران مانند سن، جنسیت، علائم بالینی، و شرایط پزشکی، مدل‌های یادگیری ماشین تلاش می‌کنند تا بیماران را به دو دسته "زنده مانده" و "زنده نمانده" طبقه‌بندی کنند.

مجموعه‌داده

این مجموعه‌داده توسط دولت مکزیک ارائه شده که شامل حجم عظیمی از اطلاعات بیماران، از جمله بیماری‌های زمینه‌ای آن‌ها به‌صورت ناشناس می‌باشد. این مجموعه‌داده از ۲۱ ویژگی منحصر به فرد و ۱,۰۴۸,۵۷۶ بیمار یکتا تشکیل شده است. در ویژگی‌های دسته‌ای، عدد ۱ به معنای "بله" و عدد ۲ به معنای "خیر" است. مقادیر ۹۷، ۹۸ و ۹۹ نشان‌دهنده داده‌های مفقود هستند. ستون‌های موجود در این مجموعه داده عبارتند از:

- sex: نشان‌دهنده جنسیت بیمار (۱ برای زن و ۲ برای مرد)
- age: سن بیمار شامل مقادیر عددی
- classification: نتیجه آزمایش کووید. مقادیر ۱-۳ به این معنی است که بیمار در درجات مختلف به کووید مبتلا شده است. ۴ یا بالاتر به این معنی است که بیمار دارای کووید نیست یا آزمایش قطعی نیست.
- patient type: نوع مراقبتی که بیمار در بخش دریافت کرده است. ۱ برای بازگشت به خانه و ۲ برای بستری شدن.
- pneumonia: بیمار قبلاً التهاب کیسه‌های هوایی داشته است یا خیر.
- pregnancy: بیمار باردار است یا خیر

- diabetes: بیمار دیابت دارد یا خیر.
- copd: بیمار به بیماری مزمن انسداد ریه مبتلا است یا خیر.
- asthma: بیمار آسم دارد یا خیر.
- inmsupr: بیمار نقص سیستم ایمنی دارد یا خیر.
- hypertension: بیمار فشار خون بالا دارد یا خیر
- cardiovascular: بیمار بیماری قبلی عروقی دارد یا خیر.
- renal chronic: بیمار بیماری مزمن کلیوی دارد یا خیر.
- other disease: بیمار مبتلا به بیماری دیگری است یا خیر.
- obesity: بیمار مبتلا به چاقی است یا خیر.
- tobacco: بیمار دخانیات مصرف می‌کند یا خیر.
- usmr: بیمار در بخش های درمانی سطح اول و دوم یا سوم تحت درمان قرار گرفته است.
- medical unit: نوع مؤسسه‌ای از نظام سلامت ملی که مراقبت را ارائه کرده است.
- intubed: بیمار به ونتیلاتور وصل شده است یا خیر.
- icu: بیمار به بخش مراقبت‌های ویژه منتقل شده است یا خیر.
- date died: اگر بیمار فوت شده باشد، تاریخ فوت آن را نشان می‌دهد. در غیر اینصورت مقدار ۹۹-۹۹-۹۹۹۹ درج شده است.

	USMER	MEDICAL UNIT	SEX	PATIENT TYPE	DATE DIED	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	...	ASTHMA	INMSUPR	HIPERTENSION	OTHER DISEASE	CARDIOVASCULAR	C
0	2	1	1	1	03/05/2020	97	1	65	2	2	...	2	2	1	2	2	
1	2	1	2	1	03/06/2020	97	1	72	97	2	...	2	2	1	2	2	
2	2	1	2	2	09/06/2020	1	2	55	97	1	...	2	2	2	2	2	
3	2	1	1	1	12/06/2020	97	2	53	2	2	...	2	2	2	2	2	
4	2	1	2	1	21/06/2020	97	2	68	97	1	...	2	2	1	2	2	

پیش پردازش

پس از وارد کردن کتابخانه‌های مورد نیاز و خواندن مجموعه‌داده، لازم است تا پیش‌پردازش‌هایی بر روی داده‌ها جهت گرفتن خروجی‌های مناسب انجام گردد. ابتدا به بررسی مقدارهای گمشده در این مجموعه‌داده می‌پردازیم. با توجه به اینکه از مقادیر NaN در داده‌ها استفاده نشده است؛ خروجی این دستور برای همه ویژگی‌ها برابر صفر خواهد بود. سپس، به بررسی مقادیر تکراری در این مجموعه‌داده می‌پردازیم. مشاهده می‌گردد که ۷۷ درصد از داده‌ها تکراری هستند. این مورد نشان‌دهنده این است که بسیاری از بیماران موجود در این مجموعه‌داده دارای شرایط کاملا مشابهی هستند. در ادامه، ویژگی‌ای به نام HAS_DIED ایجاد می‌کنیم. این ویژگی صرفا نشان‌دهنده وضعیت زنده/مرده بودن بیمار است و در ادامه مسئله به‌عنوان ویژگی هدف برای مسئله دسته‌بندی مشخص می‌گردد. لازم به ذکر است که برای ساده‌تر کردن مسئله نیز یک ویژگی باینری دیگر با نام HAS_COVID که نشان‌دهنده مبتلا بودن/نبودن بیمار است را به مجموعه‌داده خود اضافه می‌کنیم.

حال، با توجه به توضیحات مجموعه‌داده به بررسی مقادیر گمشده در ویژگی‌ها می‌پردازیم. ابتدا، میزان داده‌های گمشده را درون هر ستون مشخص می‌کنیم. اگر کمتر از نیمی از داده‌ها گمشده بود، آن‌ها را با مقداری که بیشترین تکرار را داشته (mode) پر می‌کنیم. مشاهده می‌شود که سه ویژگی 'PREGNANT', 'INTUBED', 'ICU' بیش از نیمی از مقادیرشان موجود نمی‌باشد. با بررسی‌های دقیق‌تر می‌توان متوجه شد که در این مجموعه‌داده، ناسازگاری‌هایی برای داده‌های این سه ویژگی وجود دارد. به‌عنوان مثال، برای آقایان ویژگی PREGNANT دارای مقدار گمشده بود که می‌توان با توجه به این نکته که آقایان امکان بارداری ندارند، مقدار این ویژگی را برای تمامی آن‌ها برابر False قرار داد و سایر مقادیر گمشده را نیز با توجه به عنصری که بیشتری تکرار را دارد پر کرد. همچنین داده‌ها برای ستون‌های INTUBED و ICU در افرادی که بستری نشده‌اند و به خانه بازگشتند نیز دارای مقدار گمشده بود که همانند حالت قبل آن‌ها را اصلاح می‌کنیم.

به منظور اجرای مدل‌های مختلف دسته‌بند یادگیری ماشین، ابتدا ستون AGE را با کمک MinMaxScaler به مقایس ۰ تا ۱ می‌بریم تا با بقیه ویژگی‌های‌مان در یک بازه قرار گیرد. پس از آن داده‌های کاملا تکراری را از مجموعه‌داده حذف می‌کنیم. این کار به‌جهت مقاوم‌سازی مدل و عمومیت‌سازی آن انجام می‌گیرد. لازم به ذکر است که با توجه به اینکه ۷۷ درصد داده‌ها تکراری هستند، در صورتی که این کار انجام نگیرد، پس از تقسیم مجموعه داده به دو دسته آموزش و ارزیابی، داده‌هایی که در مجموعه آموزش بودند، در مجموعه تست نیز خواهند بود و امکان ارزیابی دقیق مدل به دلیل فاش شدن داده (data leakage) وجود ندارد.

حال که داده‌های تکراری حذف شدند، ۲۰ درصد از داده‌ها را برای ارزیابی مدل و ۸۰ درصد را برای آموزش جدا می‌کنیم. با توجه به نمودار رسم شده، مشاهده می شود که داده‌های آموزش نامتوازن هستند و این مورد می‌تواند روی عملکرد مدل تاثیر منفی بگذارد.

برای حل این موضوع، با استفاده از تکنیک‌های over_smapling اقدام به افزایش داده‌های کلاس اقلیت می‌کنیم. مشاهده می‌شود که پس از انجام این کار، داده‌های دو کلاس با یکدیگر برابر شده‌اند.

بینش دریافتی از مجموعه داده

قبل از پیاده‌سازی هر گونه مدل یادگیری ماشین با تحلیل مجموعه داده، می‌توانیم به بینش‌های زیر دست یابیم:

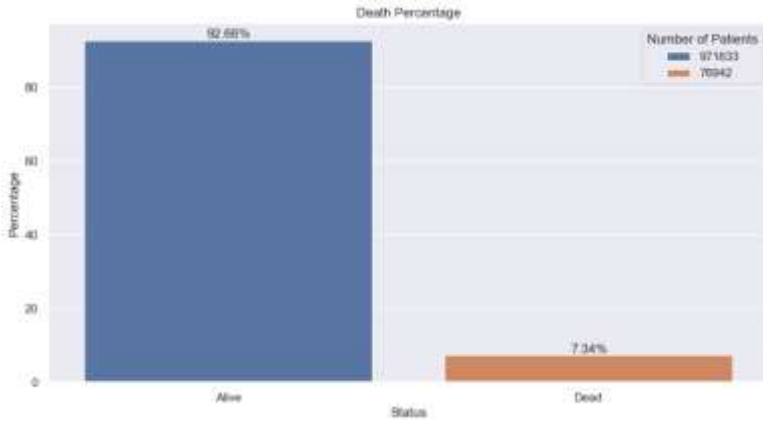
۱. **توزیع داده‌ها** :شناسایی الگوها و توزیع ویژگی‌ها مانند سن، جنسیت و مرحله تومور.

۲. **ارتباط ویژگی‌ها** :بررسی همبستگی بین متغیرها برای شناسایی ویژگی‌های مهم و تأثیرگذار.

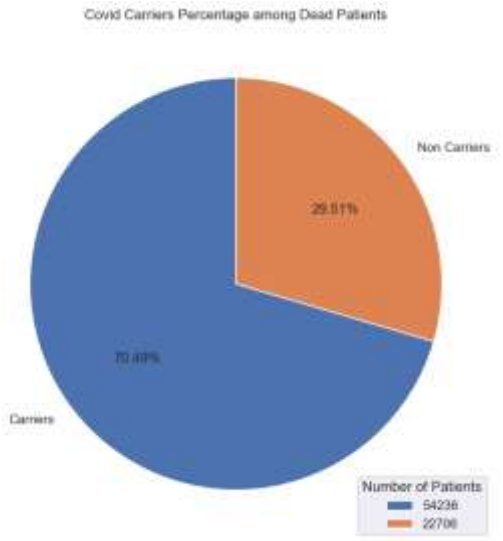
این تحلیل‌ها پایه‌ای قوی برای ساخت مدل‌های دقیق‌تر و مؤثرتر فراهم می‌کنند.

توزیع داده‌ها و ارتباط ویژگی‌ها

در این قسمت به بررسی توزیع داده‌ها با توجه به ویژگی‌های مختلف و ارتباط بین آن‌ها می‌پردازیم.

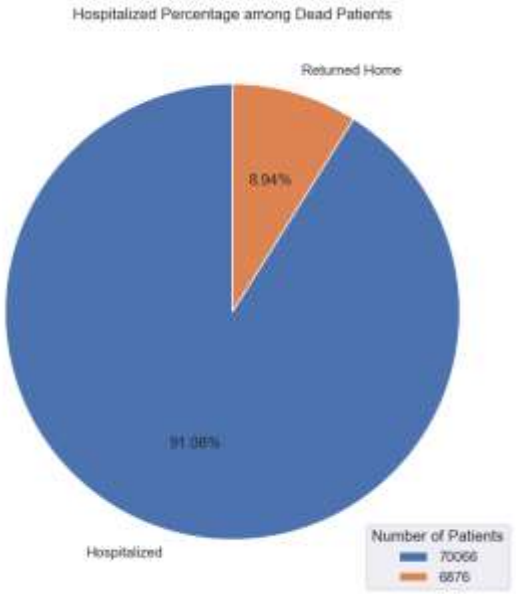


در ابتدا، به کمک نمودار میله‌ای، از نظر میزان مرگ‌ومیر، تعداد افراد فوت شده و زنده مانده در داخل مجموعه داده را نمایش می‌دهیم. مشاهده می‌شود که بیش از ۹۲ درصد از بیماران توانسته‌اند از این بیماری عبور کنند.

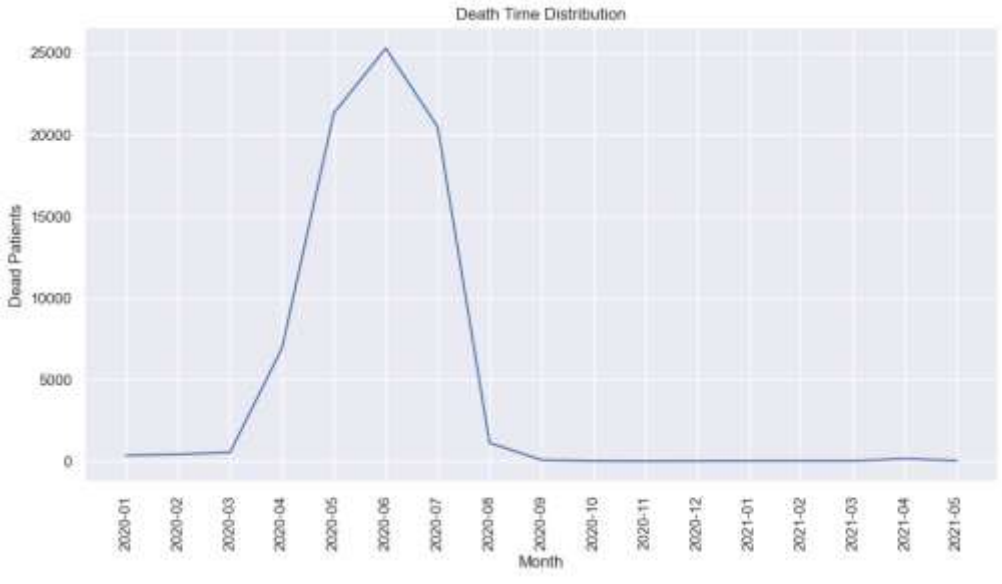


حال، در این نمودار دایره‌ای، تعداد افرادی که دارای کووید بوده‌اند و فوت شده‌اند نمایش داده شده است. می‌توان دریافت که ۷۰ درصد افراد فوت شده دارای بیماری کووید۱۹ بوده‌اند.

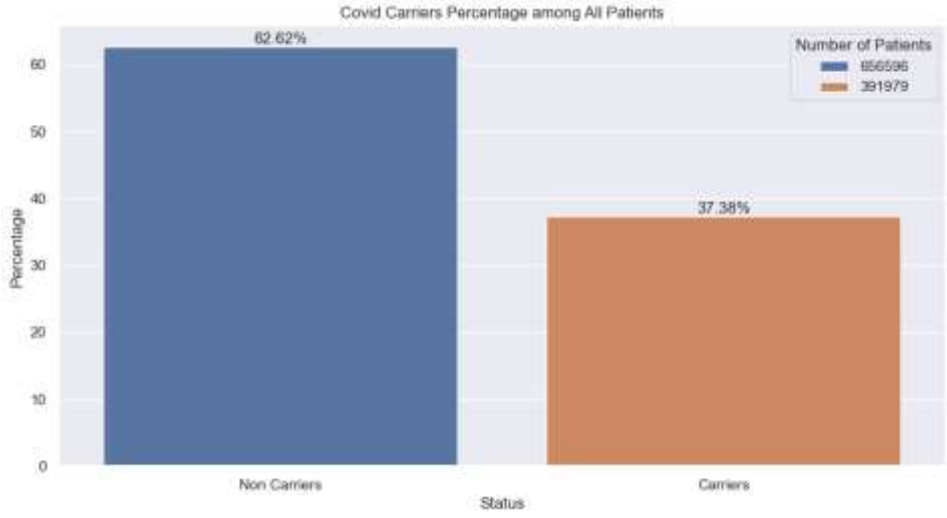
در نمودار دایره‌ای زیر، بررسی می‌کنیم چه میزان از افرادی که فوت شده‌اند، در بیمارستان بستری شده‌اند. مشاهده می‌شود که اکثریت افراد فوت شده بستری شده بودند اما تعدادی هم وجود داشته‌اند که علیرغم بستری نشدن و بازگشت به خانه فوت شده‌اند. این مورد بیانگر این است که کادر درمان در تشخیص شدت بیماری‌های این افراد دچار اشتباه شده‌اند.



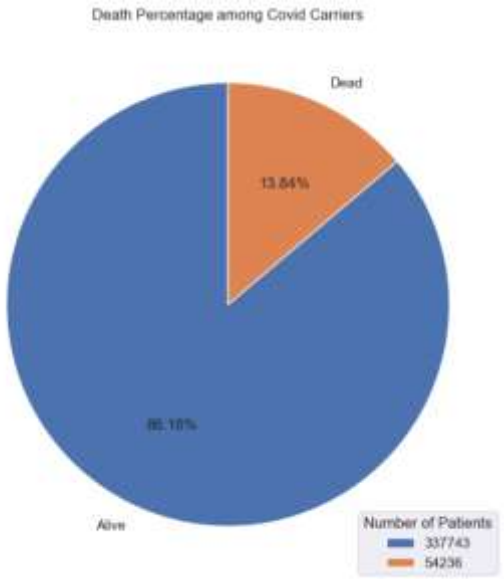
در این نمودار می‌توان توزیع زمان مرگ‌ومیر را در زمان‌های مختلف همه‌گیری بررسی کرد. می‌توان دریافت که در ماه June 2020 تعداد مرگ‌ومیر به اوج خودش رسیده بود و پس از آن با شیب نسبتا بالایی کاهش یافته است.



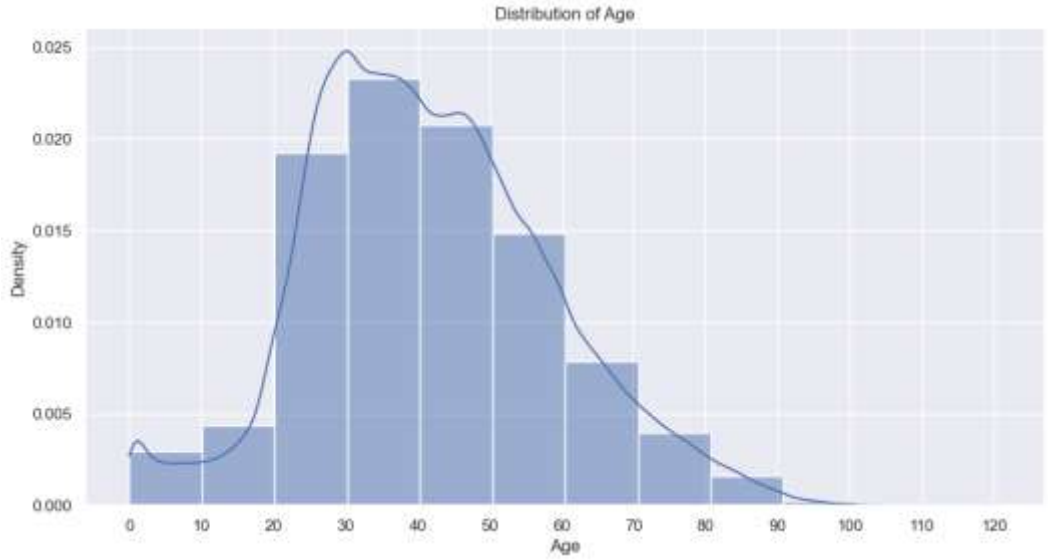
کنون، توزیع افراد دارای کووید را بررسی کرده‌ایم. مشاهده می‌شود که بیش از نیمی از افراد به کووید مبتلا نشده‌اند (ممکن است نتیجه آزمایش آن‌ها اشتباه باشد) و تقریبا ۳۷ درصد افراد دارای نتیجه آزمایش مثبت کووید هستند.



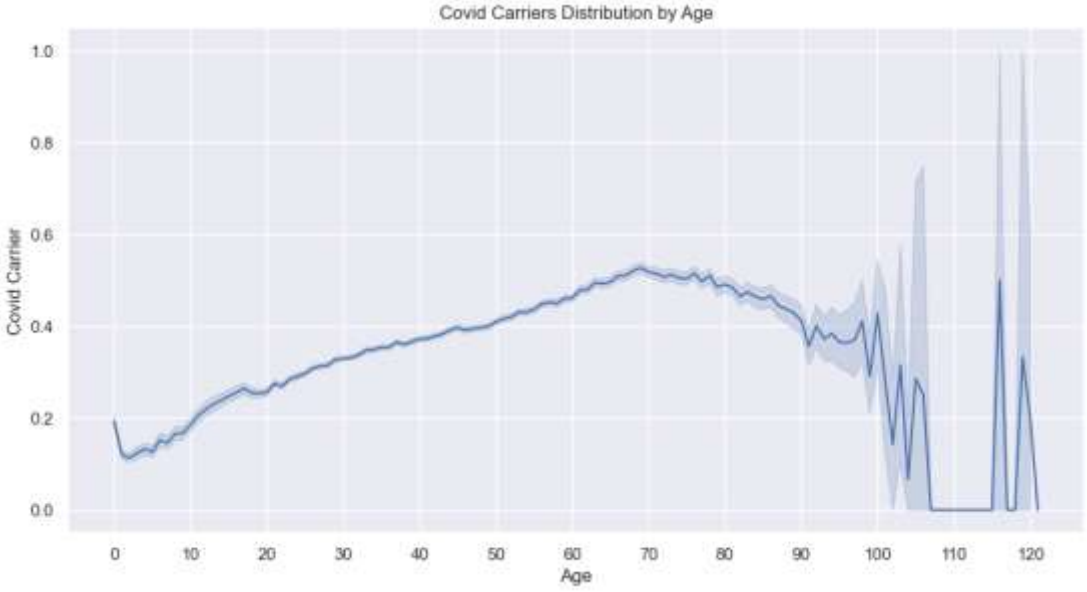
حال قصد داریم به بررسی توزیع مرگ‌ومیر در افراد مبتلا به کووید بپردازیم. مشاهده می‌کنیم که اکثر افرادی که مبتلا بودند، زنده ماندند و حدود ۱۴ درصد این افراد فوت شده‌اند.



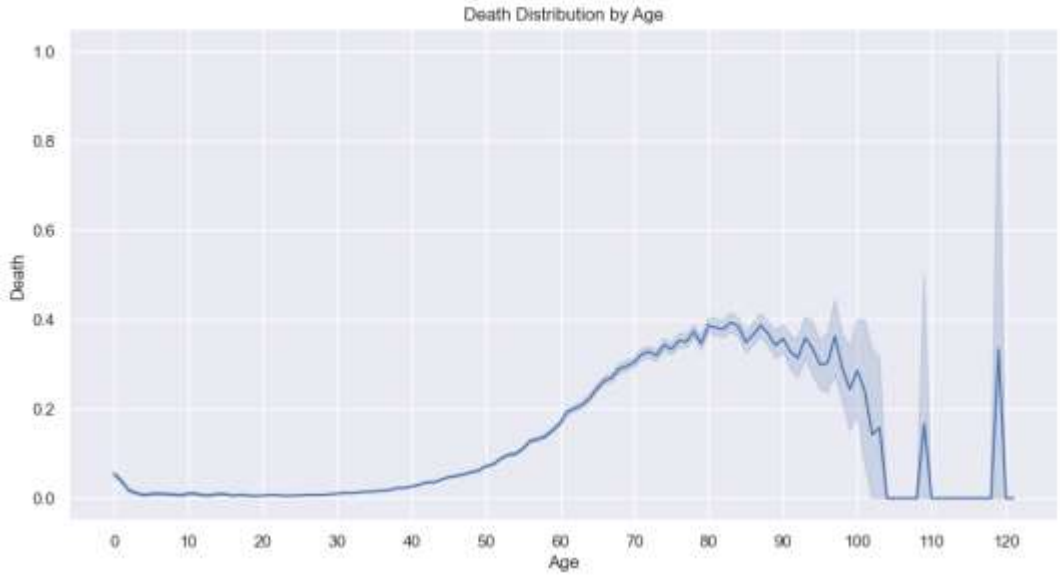
با بررسی توزیع سن در بیمارهای این مجموعه داده به کمک نمودار هیستوگرام، متوجه می‌شویم که اکثر افراد در بازه ۳۰ الی ۴۰ سال قرار گرفته‌اند.



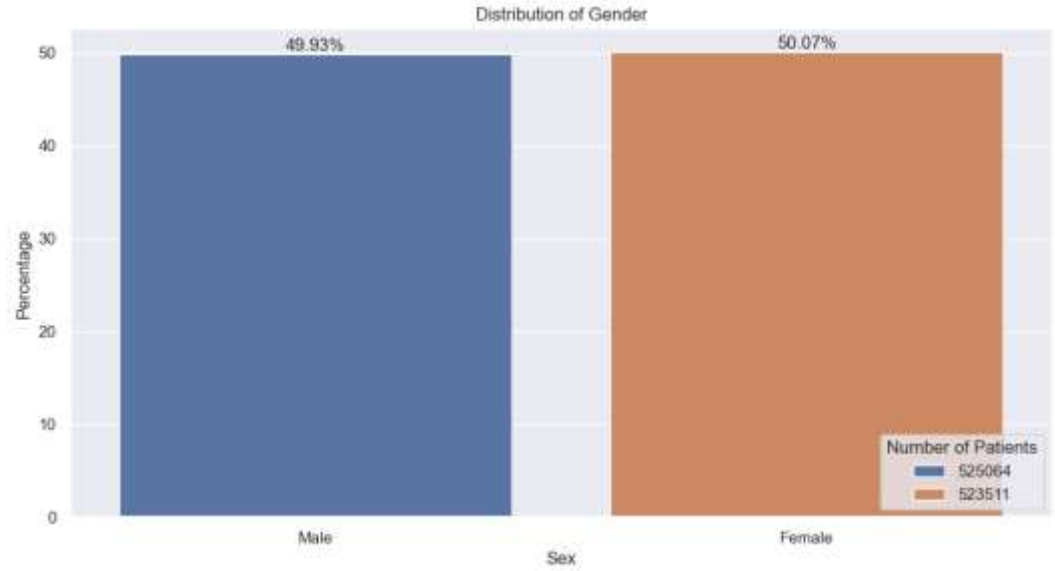
با استفاده از یک نمودار خطی، رابطه بین افزایش سن و مبتلا شدن به کووید را بررسی کرده‌ایم. از این نمودار می‌توان دریافت که هرچه سن افزایش یابد احتمال ابتلا به کووید بیشتر می‌شود. البته این قاعده تا ۷۰ سالگی برقرار است و پس از آن روند کاهشی می‌شود و در انتها الگوی چندانی ندارد.



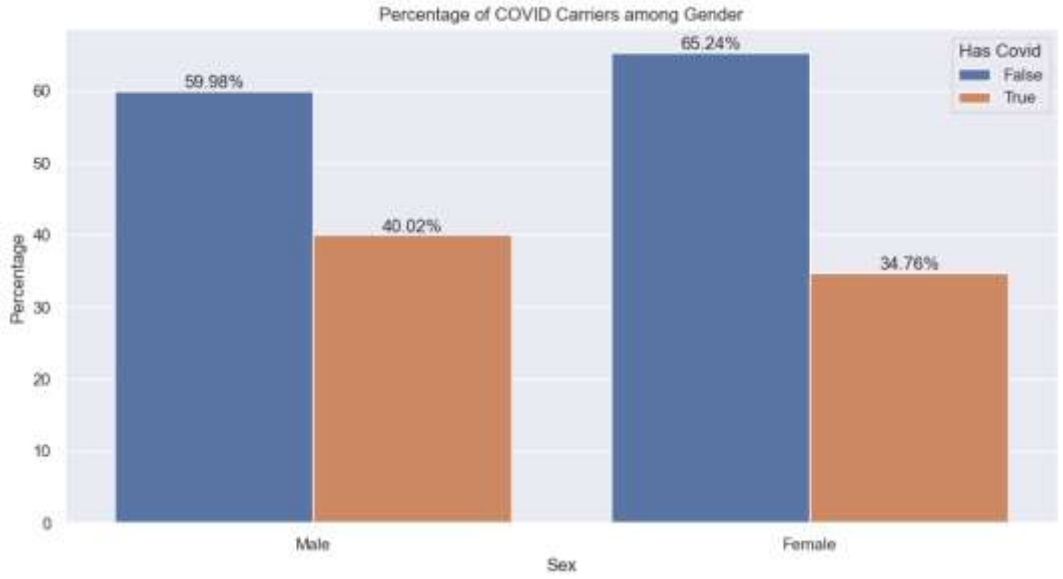
مجددا با استفاده از یک نمودار خطی، رابطه بین افزایش سن و فوت شدن را بررسی کرده‌ایم. مشاهده می‌شود که در افرادی با سن بالاتر میزان مرگ‌ومیر بیشتر است.



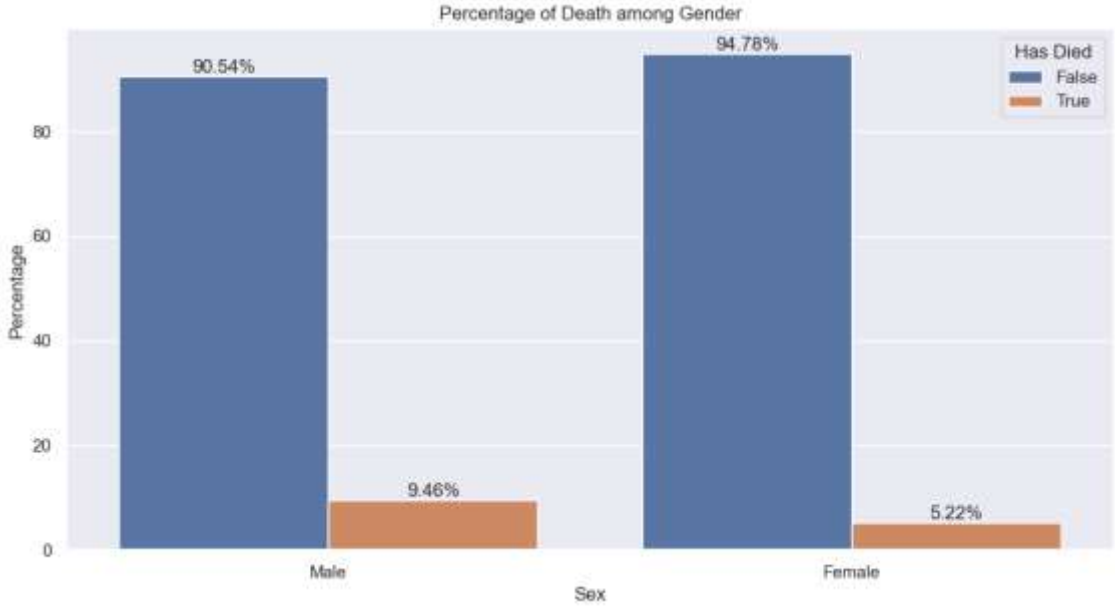
در این مرحله، به بررسی توزیع داده‌های جنسیت می‌پردازیم. می‌توانیم مشاهده کنیم که جنسیت افراد تقریبا به صورت متوازن در این مجموعه‌داده پخش شده است.



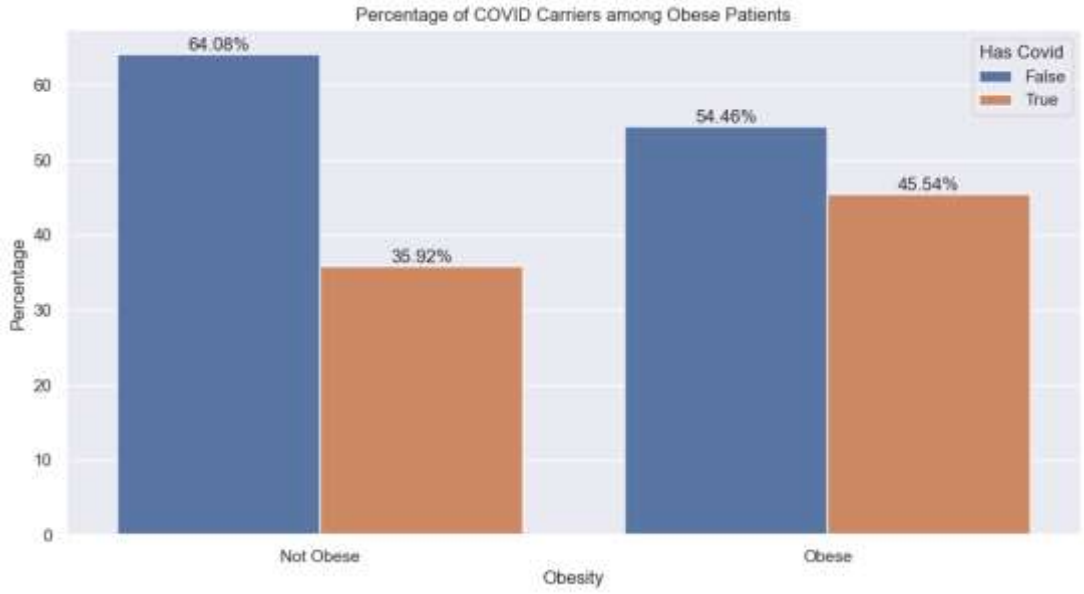
حال قصد داریم به بررسی تاثیر جنسیت بر روی ابتلا به کووید بپردازیم. مشاهده می‌کنیم که ابتلا در آقایان کمی بیشتر از خانم‌ها بوده است. اما با توجه به اینکه اختلاف کمی میان این دو وجود دارد، می‌توان دریافت که جنسیت اثر خاصی بر روی ابتلا به کووید ندارد.



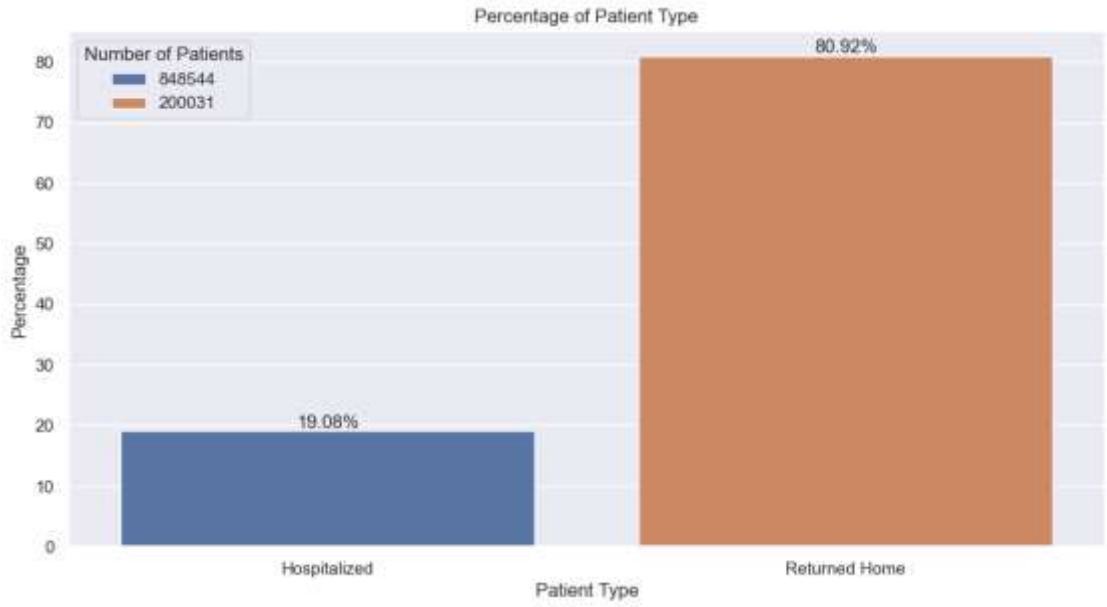
در این مرحله، همانند مرحله قبل، قصد داریم تا اثر جنسیت را بر روی مرگ‌ومیر بررسی کنیم. همانند قبل نتیجه‌گیری می‌شود که مرگ‌ومیر در خانم‌ها کمی کمتر از آقایان بوده و با توجه به فاصله کم این دو مورد، می‌توان گفت جنسیت نیز اثر چندانی بر روی مرگ‌ومیر ندارد.



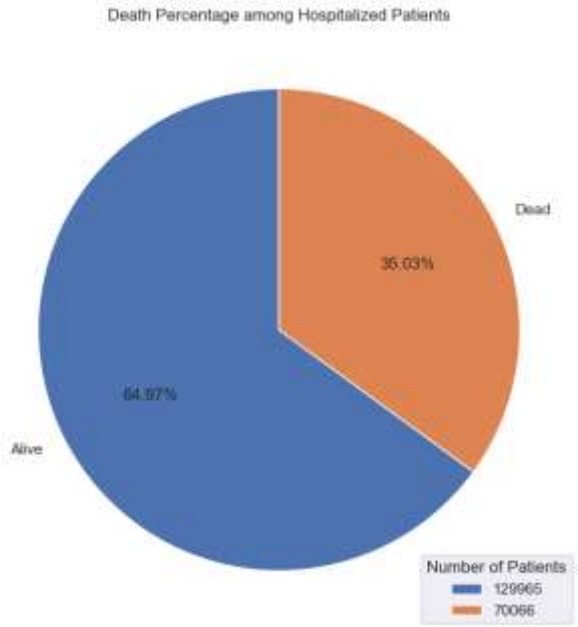
حال، قصد داریم تا اثر چاقی را بر روی ابتلا به کووید بررسی کنیم. مشاهده می‌کنیم که در بین افرادی که چاق هستند، ابتلا به کووید نسبت به افرادی که چاق نبوده‌اند، بیشتر بوده است. این می‌تواند بیانگر این موضوع باشد که چاقی ممکن است یکی از عوامل موثر در ابتلا به این بیماری باشد.



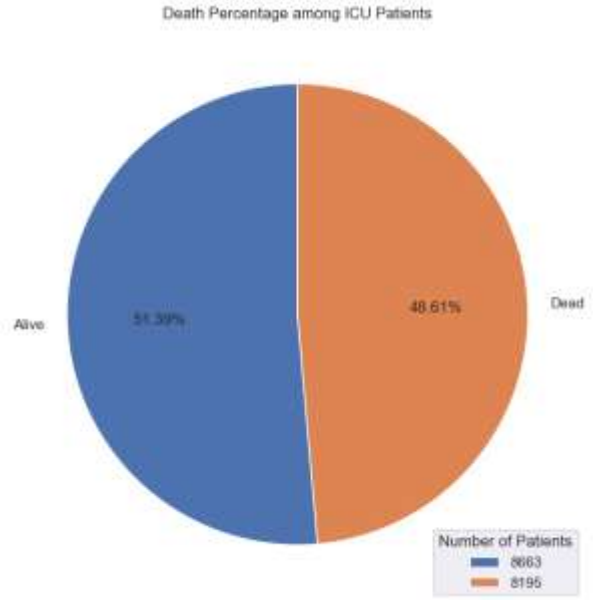
در این نمودار، بررسی می‌کنیم که چه تعداد از بیمارن در بیمارستان بستری و چه تعداد به خانه بازگشتند. مشاهده می‌شود که نزدیک ۸۱ درصد از بیمارن نیاز به مراقبت ویژه‌ای نداشتند و به خانه بازگشتند.



در نمودار دایره‌ای زیر، از بین افرادی که بستری شده‌اند؛ بررسی می‌کنیم که چه تعداد فوت شده و چه تعداد زنده ماندند. مشاهده می‌شود که تقریبا ۳۵ درصد از افراد بستری شده فوت شدند.



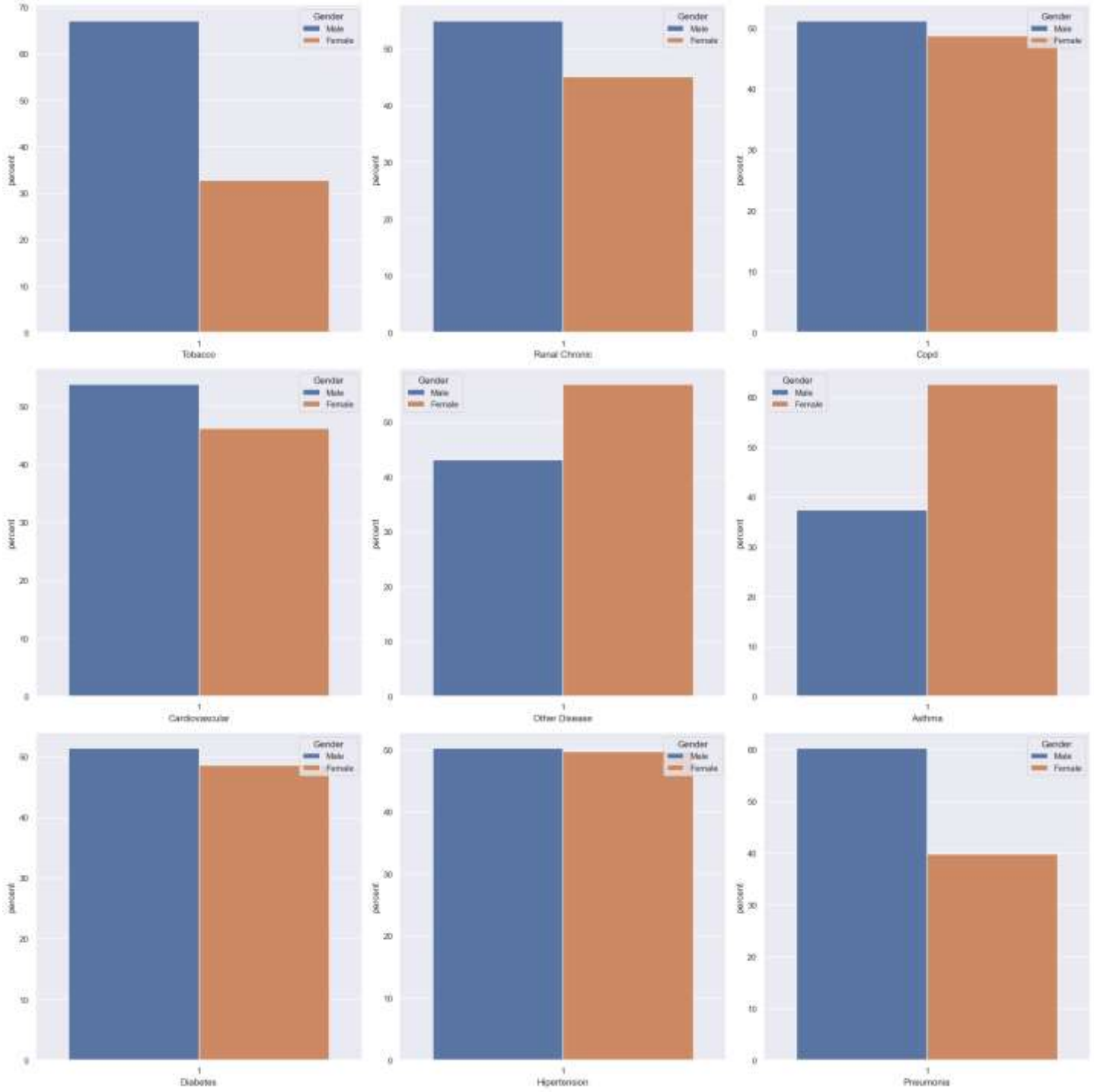
در این نمودار نیز بررسی کرده‌ایم که چه تعداد از افرادی که در بخش مراقبت‌های ویژه بستری شده‌اند، فوت شدند. در اینجا تقریبا نیمی از افراد بستری شده در ICU فوت شده‌اند.



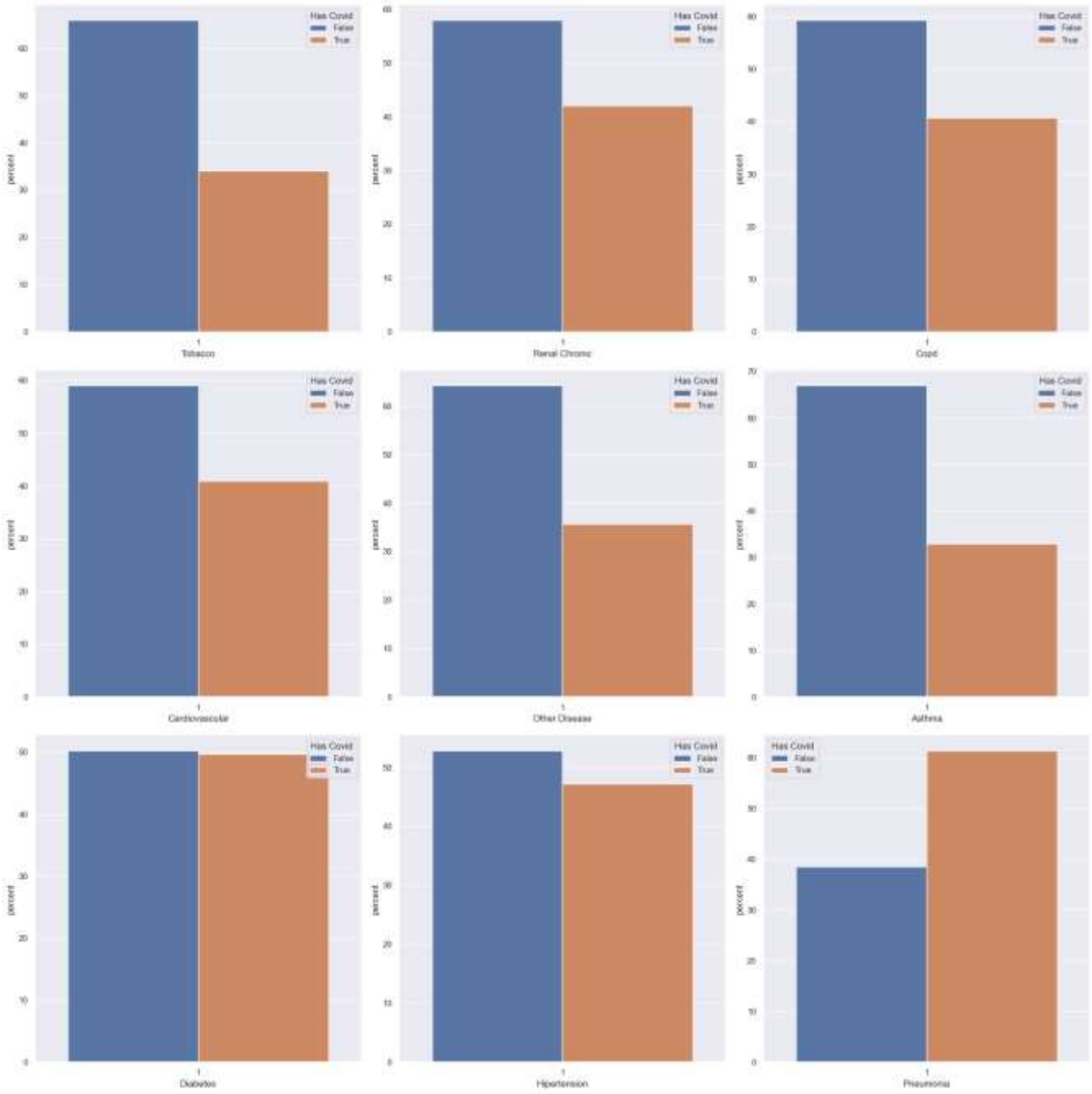
در سری نمودار زیر قصد داریم تا وجود داشتن ویژگی‌های مختلف را بر اساس جنسیت مقایسه کنیم.

- مصرف دخانیات: مشاهده می‌شود که آقایان مصرف بیشتری از دخانیات و فاصله زیادی نیز با خانم‌ها دارند.

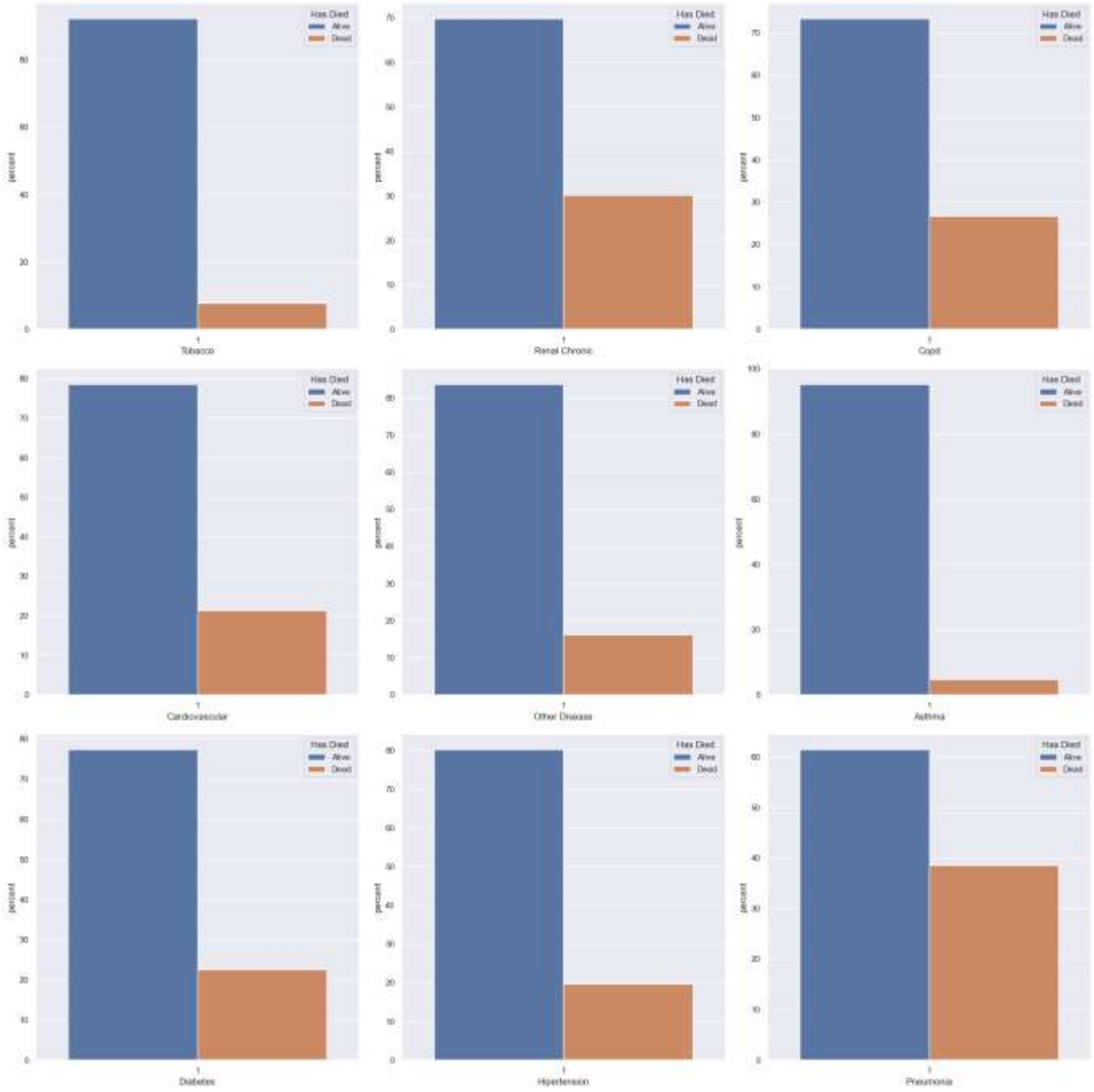
- بیماری مزمن کلیوی: مشاهده می‌شود که در بین افراد مبتلا به این بیماری، آقایان بیشتر از خانم‌ها هستند. البته لازم به ذکر است که اختلاف تنها حدود ۱۰ درصد است.
- بیماری مزمن انسداد ریه: مشاهده می‌شود که در بین افراد مبتلا به این بیماری، تقریبا توزیع جنسیت برابر است.
- بیماری قلبی یا عروقی: همانند حالت بیماری کلیوی، در این نمودار هم مشاهده می‌شود که در بین افراد مبتلا به این بیماری، آقایان بیشتر از خانم‌ها هستند. البته لازم به ذکر است که اختلاف تنها حدود ۱۰ درصد است.
- سایر بیماری‌ها: در بین افرادی که به سایر بیماری‌ها مبتلا بودند، تعداد خانم‌ها تقریبا ۱۵ درصد بیش‌تر از آقایان هست.
- آسم: در بین افراد مبتلا به آسم، خانم‌ها با اختلاف نسبتا بالایی در صدر قرار دارند.
- دیابت: در بین افراد مبتلا به دیابت، می‌توان مشاهده کرد که توزیع جنسیت تقریبا برابر است.
- فشار خون بالا: مجددا در این دسته هم افراد مبتلا به فشار خون بالا داری توزیع جنسیت برابری هستند.
- التهاب کیسه‌های هوایی: در این دسته مشاهده می‌شود که اکثر افراد مبتلا به این بیماری آقا هستند. (با بیش از ۲۰ درصد اختلاف)


Gender
Male
Female
percent
1
Tobacco
Renal Chronic
Copd
Cardiovascular
Other Disease
Asthma
Diabetes
Hypertension
Pneumonia

در سری نمودارهای زیر قصد داریم تا وجود داشتن ویژگی‌های مختلف را بر اساس مبتلا بودن به کووید مقایسه کنیم. تقریبا می‌توانیم از این نمودارها متوجه شویم که وجود اکثر این علائم باعث افزایش مبتلایان به کووید در آن دسته نشده است. به جز دیابت و التهاب کیسه‌های هوایی که در این دو حالت افرادی که دارای این ویژگی‌ها بودند بیش از نیمی از آن‌ها به کووید مبتلا شدند.



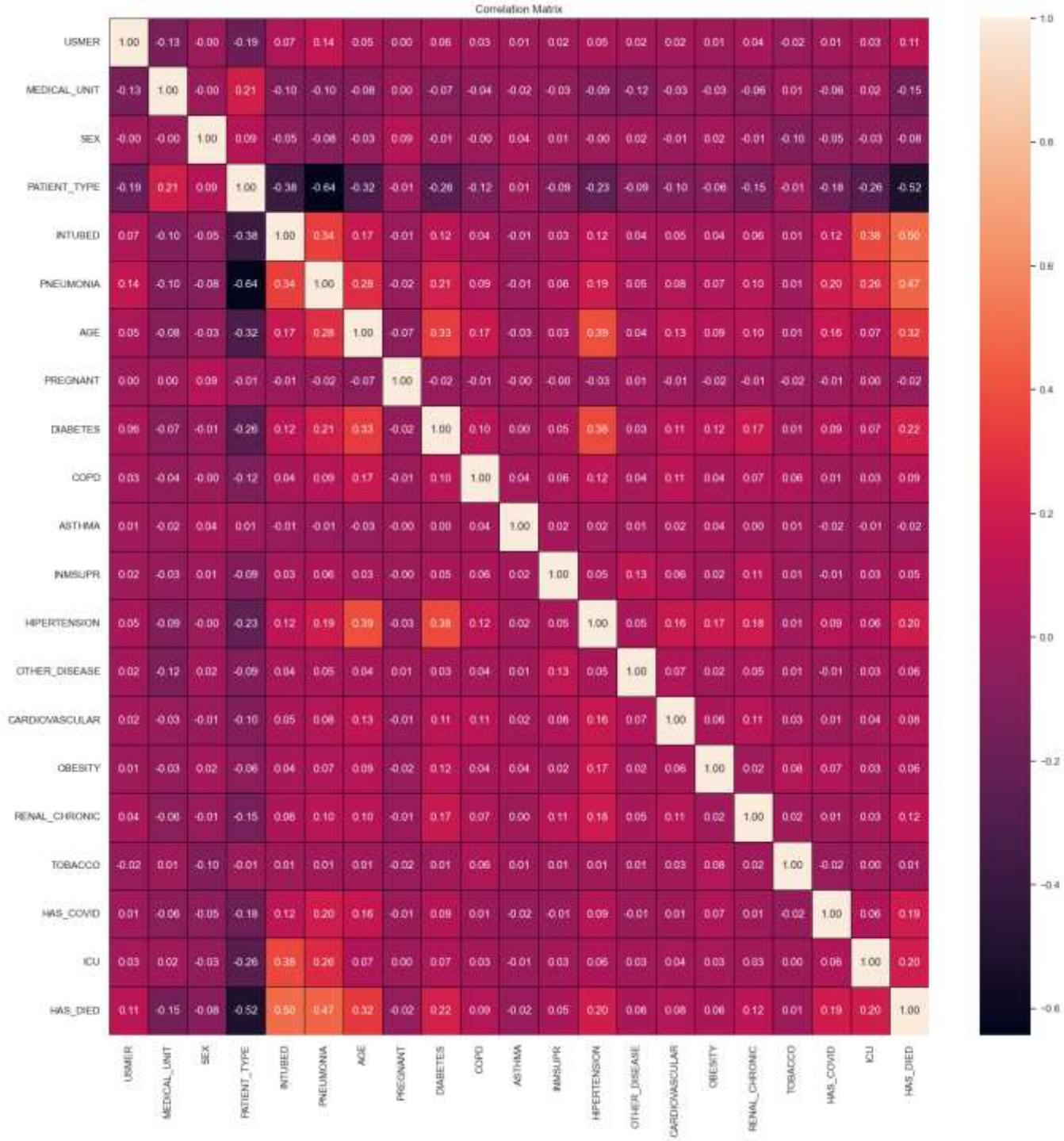
در سری نمودارهای زیر قصد داریم تا وجود داشتن ویژگی‌های مختلف را بر اساس فوت شدن/زنده ماندن مقایسه کنیم. می‌توانیم از این نمودارها در بیابیم که التهاب کیسه‌های هوایی و بیماری‌های قلبی و عروقی بر فوت شدن افراد اثر بیشتری دارند.



ماتریس همبستگی

ماتریس همبستگی جدولی مربعی است که در آن ضرایب همبستگی پیرسون بین هر دو متغیر در یک مجموعه داده نشان داده می‌شود. هرچه ضریب همبستگی بین دو متغیر به ۱ نزدیک‌تر باشد، نشان‌دهنده همبستگی قوی‌تر و مثبت‌تر بین آن دو است. به عبارت دیگر، مقادیر مثبت نشان می‌دهند که با افزایش یک متغیر، به احتمال زیاد متغیر دیگر نیز افزایش می‌یابد. در مقابل، مقادیر منفی نشان‌دهنده همبستگی معکوس است، به این معنی که با افزایش یک متغیر، به احتمال زیاد متغیر دیگر کاهش می‌یابد. با توجه به این ماتریس می‌توانیم نکات زیر را استخراج کنیم:

- میان فوت شدن فرد و ویژگی‌های pneumonia، intubed و patient type همبستگی بالایی وجود دارد. این نشان دهنده این است که التهاب کیسه هوایی، وصل شدن به ونتیلاتور و همچنین بستری شدن در بیمارستان رابطه مستقیم و زیادی با مرگ شخص دارد. پس از این ستون‌ها، به‌ترتیب سن، دیابت، بستری شدن در ICU، فشارخون بالا و داشتن کووید رابطه تقریبا بالایی با مرگ فرد دارند.
- میان سن یک فرد و فشار خون بالا، دیابت، فوت شخص همبستگی مثبت و بالایی دیده می‌شود. این نشان‌دهنده این است که با افزایش سن، احتمال وجود دیابت و فشار خون بالا و همچنین احتمال فوت شخص بالاتر می‌رود. همچنین می‌توان از همبستگی منفی سن و بستری شدن در بیمارستان به این نتیجه رسید که هرچه سن بالاتر می‌رود احتمال بستری در بیمارستان بالاتر است.
- همچنین از رابطه بستری شدن و التهاب کیسه‌های هوایی می‌توان این برداشت را کرد که این مورد یکی از عوامل اصلی بستری شدن بیماران است.


Correlation Matrix
USMER 1.00 -0.13 -0.00 -0.19 0.07 0.14 0.05 0.00 0.06 0.03 0.01 0.02 0.05 0.02 0.02 0.01 0.04 -0.02 0.01 0.03 0.11
MEDICAL_UNIT -0.13 1.00 -0.00 0.21 -0.10 -0.10 -0.08 0.00 -0.07 -0.04 -0.02 -0.03 -0.09 -0.12 -0.03 -0.03 -0.06 0.01 -0.06 0.02 -0.15
SEX -0.00 -0.00 1.00 0.09 -0.05 -0.08 -0.03 0.09 -0.01 -0.00 0.04 0.01 -0.00 0.02 -0.01 0.02 -0.01 -0.10 -0.05 -0.03 -0.08
PATIENT_TYPE -0.19 0.21 0.09 1.00 -0.38 -0.64 -0.32 -0.01 -0.26 -0.12 0.01 -0.09 -0.23 -0.09 -0.10 -0.06 -0.15 -0.01 -0.18 -0.26 -0.52
INTUBED 0.07 -0.10 -0.05 -0.38 1.00 0.34 0.17 -0.01 0.12 0.04 -0.01 0.03 0.12 0.04 0.05 0.04 0.06 0.01 0.12 0.38 0.00
PNEUMONIA 0.14 -0.10 -0.08 -0.64 0.34 1.00 0.28 -0.02 0.21 0.09 -0.01 0.06 0.19 0.05 0.08 0.07 0.10 0.01 0.20 0.26 0.47
AGE 0.05 -0.08 -0.03 -0.32 0.17 0.28 1.00 -0.07 0.33 0.17 -0.03 0.03 0.39 0.04 0.13 0.09 0.10 0.01 0.16 0.07 0.32
PREGNANT 0.00 0.00 0.09 -0.01 -0.01 -0.02 -0.07 1.00 -0.02 -0.01 -0.00 -0.00 -0.03 0.01 -0.01 -0.02 -0.01 -0.02 -0.01 0.00 -0.02
DIABETES 0.06 -0.07 -0.01 -0.26 0.12 0.21 0.33 -0.02 1.00 0.10 0.00 0.05 0.36 0.03 0.11 0.12 0.17 0.01 0.09 0.07 0.22
COPD 0.03 -0.04 -0.00 -0.12 0.04 0.09 0.17 -0.01 0.10 1.00 0.04 0.06 0.12 0.04 0.11 0.04 0.07 0.06 0.01 0.03 0.09
ASTHMA 0.01 -0.02 0.04 0.01 -0.01 -0.01 -0.03 -0.00 0.00 0.04 1.00 0.02 0.02 0.01 0.02 0.04 0.00 0.01 -0.02 -0.01 -0.02
INMSUPR 0.02 -0.03 0.01 -0.09 0.03 0.06 0.03 -0.00 0.05 0.06 0.02 1.00 0.05 0.13 0.06 0.02 0.11 0.01 -0.01 0.03 0.05
HIPERTENSION 0.05 -0.09 -0.00 -0.23 0.12 0.19 0.39 -0.03 0.36 0.12 0.02 0.05 1.00 0.05 0.16 0.17 0.18 0.01 0.09 0.06 0.20
OTHER_DISEASE 0.02 -0.12 0.02 -0.09 0.04 0.05 0.04 0.01 0.03 0.04 0.01 0.13 0.05 1.00 0.07 0.02 0.05 0.01 -0.01 0.03 0.06
CARDIOVASCULAR 0.02 -0.03 -0.01 -0.10 0.05 0.08 0.13 -0.01 0.11 0.11 0.02 0.06 0.16 0.07 1.00 0.06 0.11 0.03 0.01 0.04 0.08
OBESITY 0.01 -0.03 0.02 -0.06 0.04 0.07 0.09 -0.02 0.12 0.04 0.04 0.02 0.17 0.02 0.06 1.00 0.02 0.08 0.07 0.03 0.06
RENAL_CHRONIC 0.04 -0.06 -0.01 -0.15 0.06 0.10 0.10 -0.01 0.17 0.07 0.00 0.11 0.16 0.05 0.11 0.02 1.00 0.02 0.01 0.03 0.12
TOBACCO -0.02 0.01 -0.10 -0.01 0.01 0.01 0.01 -0.02 0.01 0.06 0.01 0.01 0.01 0.01 0.03 0.08 0.02 1.00 -0.02 0.00 0.01
HAS_COVID 0.01 -0.06 -0.05 -0.18 0.12 0.20 0.16 -0.01 0.09 0.01 -0.02 -0.01 0.09 -0.01 0.01 0.07 0.01 -0.02 1.00 0.06 0.19
ICU 0.03 0.02 -0.03 -0.26 0.38 0.26 0.07 0.00 0.07 0.03 -0.01 0.03 0.06 0.03 0.04 0.03 0.03 0.00 0.06 1.00 0.20
HAS_DIED 0.11 -0.15 -0.08 -0.52 0.50 0.47 0.32 -0.02 0.22 0.09 -0.02 0.05 0.20 0.06 0.08 0.06 0.12 0.01 0.19 0.20 1.00
USMER
MEDICAL_UNIT
SEX
PATIENT_TYPE
INTUBED
PNEUMONIA
AGE
PREGNANT
DIABETES
COPD
ASTHMA
INMSUPR
HIPERTENSION
OTHER_DISEASE
CARDIOVASCULAR
OBESITY
RENAL_CHRONIC
TOBACCO
HAS_COVID
ICU
HAS_DIED
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

پیاده‌سازی مدل‌های یادگیری ماشین

در این بخش با استفاده از کتابخانه Scikit-learn در زبان پایتون به پیاده‌سازی انواع مدل‌های یادگیری ماشین از جمله رگرسیون لاجستیک، بیز ساده‌لوحانه، درخت تصمیم، بردارهای پشتیبان خطی، جنگل تصادفی، تقویت گرادیان، شبکه پرسپترون عصبی چندلایه و k نزدیک‌ترین همسایه پرداخته و عملکرد هر یک از این مدل‌ها را با استفاده از معیارهای مختلف در مقایسه با یکدیگر بررسی می‌کنیم.

معیارهای ارزیابی

پس از آموزش مدل لازم است عملکرد آن را بسنجیم. این کار را می‌توان با استفاده از معیارهای مختلف مانند دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 انجام داد.

- **دقت (Precision)**: دقت، نسبت نمونه‌های صحیح مثبت به کل نمونه‌های پیش‌بینی‌شده مثبت است. به عبارت دیگر، دقت نشان می‌دهد که از میان همه مواردی که مدل آنها را مثبت پیش‌بینی کرده است، چه تعداد واقعاً مثبت بوده‌اند. دقت بالا نشان می‌دهد که مدل دارای تعداد کمی مثبت کاذب است.

$$Precision = \frac{TP}{TP + FP}$$

- **بازیابی یا حساسیت (Recall)**: بازیابی، نسبت نمونه‌های صحیح مثبت به کل نمونه‌های واقعی مثبت است. این معیار نشان می‌دهد که مدل چه تعداد از نمونه‌های واقعی مثبت را به درستی شناسایی کرده است. بازیابی بالا نشان می‌دهد که مدل دارای تعداد کمی منفی کاذب است.

$$Recall = \frac{TP}{TP + FN}$$

- **امتیاز F1**: امتیاز F1 میانگین هارمونیک دقت و بازیابی است. این یک معیار متوازن است که هر دو مثبت کاذب و منفی کاذب را در نظر می‌گیرد. امتیاز F1 زمانی مفید است که می‌خواهیم یک معیار واحد برای ارزیابی عملکرد مدل داشته باشیم. این معیار به ویژه زمانی مفید است که بین مثبت کاذب و منفی کاذب تعادل وجود داشته باشد.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

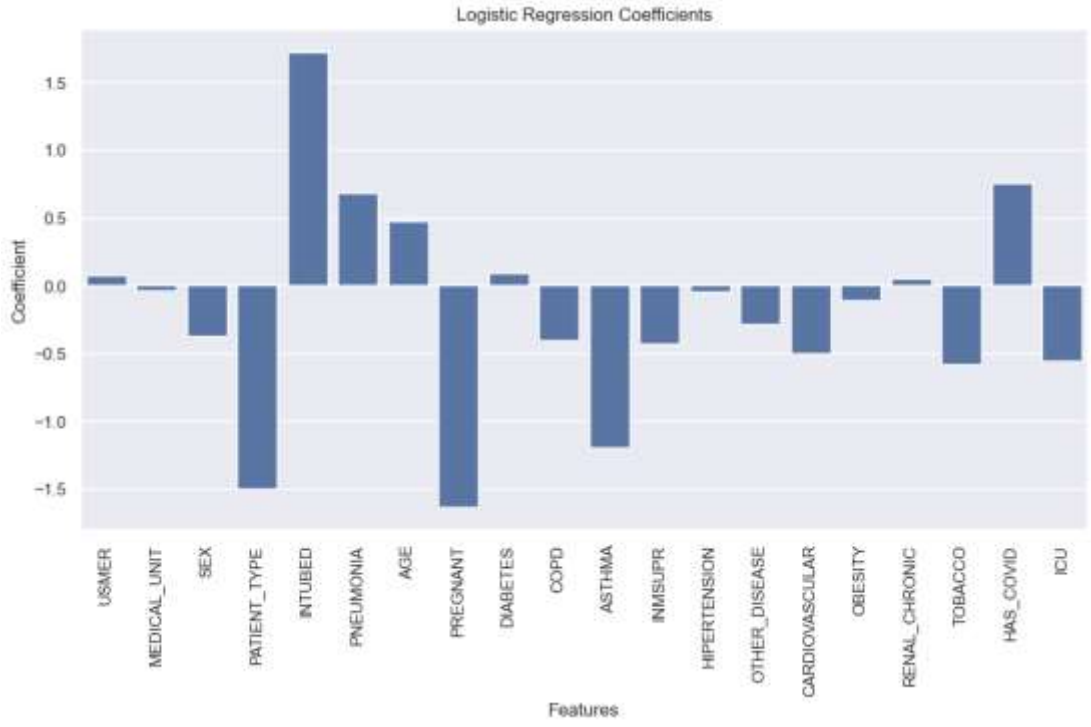
رگرسیون لاجستیک

رگرسیون لاجستیک یک الگوریتم یادگیری ماشین نظارت‌شده است که برای انجام وظایف دسته‌بندی دوتایی استفاده می‌شود. این روش از تابع سیگموئید برای تبدیل یک ترکیب خطی از ویژگی‌های ورودی به یک مقدار احتمالی بین ۰ و ۱ استفاده می‌کند. این تکنیک

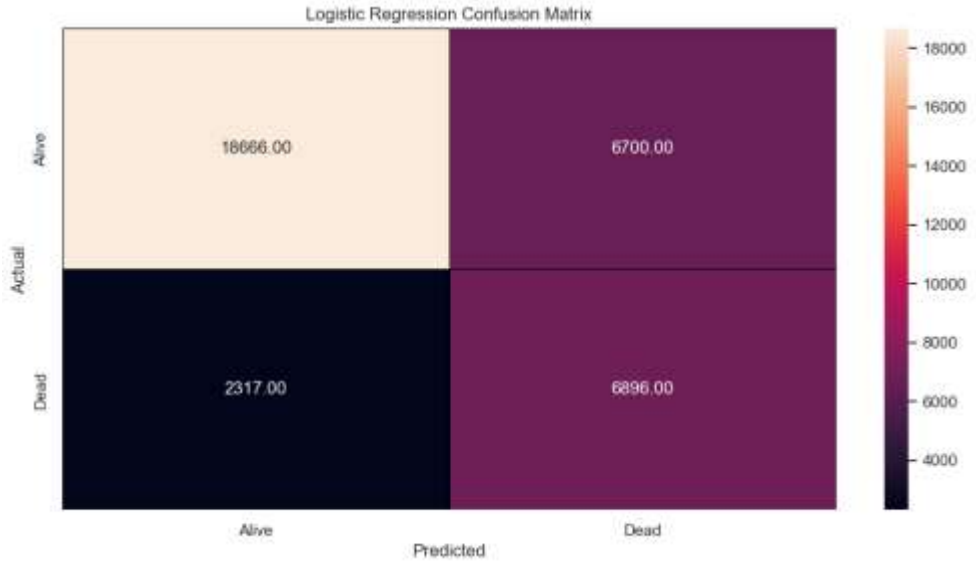
مزایایی مانند سادگی در پیاده‌سازی و تفسیر، توانایی مدیریت متغیرهای مستقل چندگانه و دسته‌بندی داده‌ها به کلاس‌های گسسته را ارائه می‌دهد. با این حال، فرضیات خطی و نیاز به نمونه‌های بزرگتر برای دقت ممکن است محدودیت‌هایی ایجاد کند. با استفاده از ماژول رگرسیون لاجستیک در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل رگرسیون لاجستیک معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Logistic Regression Classification Report:
Training Time: 2.8545s | Prediction Time: 0.0055s

              precision    recall  f1-score   support

       False       0.89      0.74      0.81     25366
        True       0.51      0.75      0.60      9213

    accuracy                           0.74     34579
   macro avg       0.70      0.74      0.71     34579
weighted avg       0.79      0.74      0.75     34579
```

همچنین در نمودار زیر می‌توانیم ضرایب آموزش‌دیده‌شده توسط الگوریتم برای هر ویژگی را مشاهده می‌کنیم. اندازه این ضریب‌ها می‌تواند نشان‌دهنده اهمیت آن باشد.



همچنین ماتریس درهم‌ریختگی این مدل نیز به شرح زیر می‌باشد. نکته قابل توجه این هست که میزان FN در مدل ما تا حد امکان پایین باشد. زیرا این مورد نشان‌دهنده این است که مدل پیش‌بینی کرده است که بیمار زنده می‌ماند اما در عمل بیمار می‌میرد و این مورد می‌تواند خطرات جدی‌ای داشته باشد.



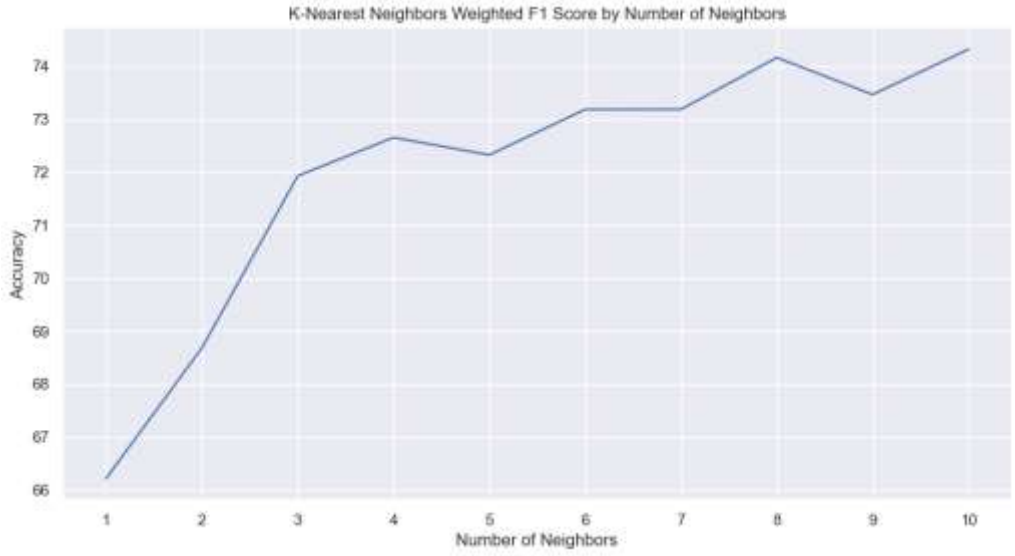
K نزدیک‌ترین همسایه

یک الگوریتم یادگیری نظارت‌نشده است که دسته‌بندی را بر اساس داده‌های نزدیک‌ترین همسایه‌ها انجام می‌دهد. این الگوریتم تعداد مشخصی از نزدیک‌ترین همسایه‌ها را در نظر می‌گیرد و دسته‌بندی را بر اساس رای‌گیری ساده‌ی اکثریت انجام می‌دهد. این روش به کاربر اجازه می‌دهد تا تعداد همسایگان (k) را مشخص کند که بر میزان سرکوب نویز و تمایز مرزهای طبقه‌بندی تاثیر می‌گذارد. این الگوریتم دسته‌بند در طیف وسیعی از کاربردهای طبقه‌بندی استفاده می‌شود، مانند طبقه‌بندی یک ستاره جدید بر اساس ویژگی‌های آن. با استفاده از ماژول k نزدیک‌ترین همسایه در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیازF1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل k نزدیک‌ترین همسایه معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

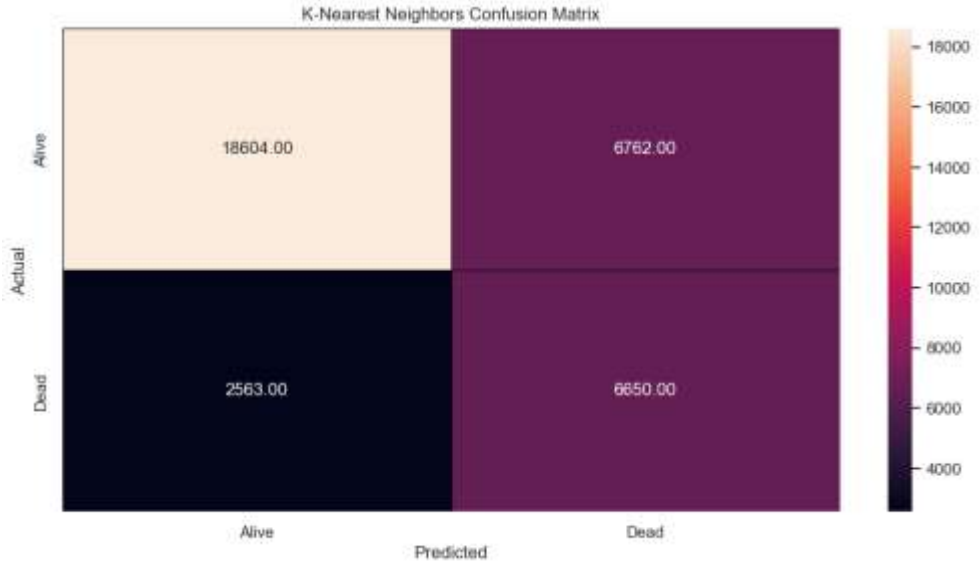
K-Nearest Neighbors Classification Report:
Training Time: 0.037s | Prediction Time: 4.685s

	precision	recall	f1-score	support
False	0.87	0.71	0.78	25366
True	0.47	0.70	0.56	9213
accuracy			0.71	34579
macro avg	0.67	0.71	0.67	34579
weighted avg	0.76	0.71	0.72	34579

برای اینکه اثر k را در عملکرد مدل بررسی کنیم؛ این الگوریتم را به ازای kهای مختلف اجرا کردیم و بر اساس f1-score عملکردشان را مقایسه کردیم.



همچنین ماتریس درهم‌ریختگی این مدل نیز به شرح زیر می‌باشد. نکته قابل توجه این هست که میزان FN در مدل ما تا حد امکان پایین باشد. زیرا این مورد نشان‌دهنده این است که مدل پیش‌بینی کرده است که بیمار زنده می‌ماند اما در عمل بیمار می‌میرد و این مورد می‌تواند خطرات جدی‌ای داشته باشد.

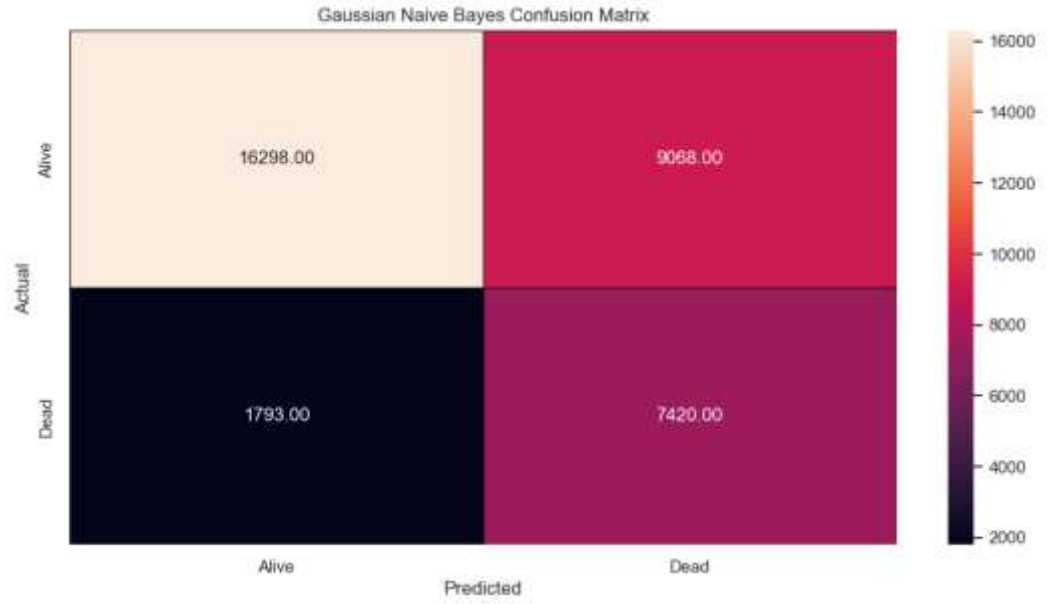


بیز ساده‌لوحانه

یک الگوریتم دسته‌بندی محبوب و قدرتمند است که بر اساس تئوری احتمال بیز و فرض توزیع نرمال ساخته شده است. این روش فرض می‌کند که ویژگی‌های ورودی از توزیع گوسی پیروی می‌کنند و از آن برای تخمین احتمال عضویت در یک کلاس خاص استفاده می‌کند. این الگوریتم ساده و کارآمد است و برای داده‌هایی با توزیع نرمال یا نزدیک به نرمال بسیار مناسب است. طبقه‌بندی بیز ساده گاوسی در طیف وسیعی از کاربردهای دنیای واقعی استفاده می‌شود، از جمله پردازش زبان طبیعی، تشخیص چهره و فیلتر اسپم ایمیل. این الگوریتم انعطاف‌پذیر است و می‌تواند با انواع مختلفی از داده‌ها سازگار شود. با استفاده از ماژول بیز ساده‌لوحانه در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیازF1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل بیز ساده‌لوحانه معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Gaussian Naive Bayes Classification Report:
Training Time: 0.090s | Prediction Time: 0.013s

              precision    recall  f1-score   support

       False       0.90      0.64      0.75     25366
        True       0.45      0.81      0.58      9213

    accuracy                           0.69     34579
   macro avg       0.68      0.72      0.66     34579
weighted avg       0.78      0.69      0.70     34579
```

همچنین ماتریس درهم‌ریختگی این مدل نیز به شکل زیر می‌باشد.



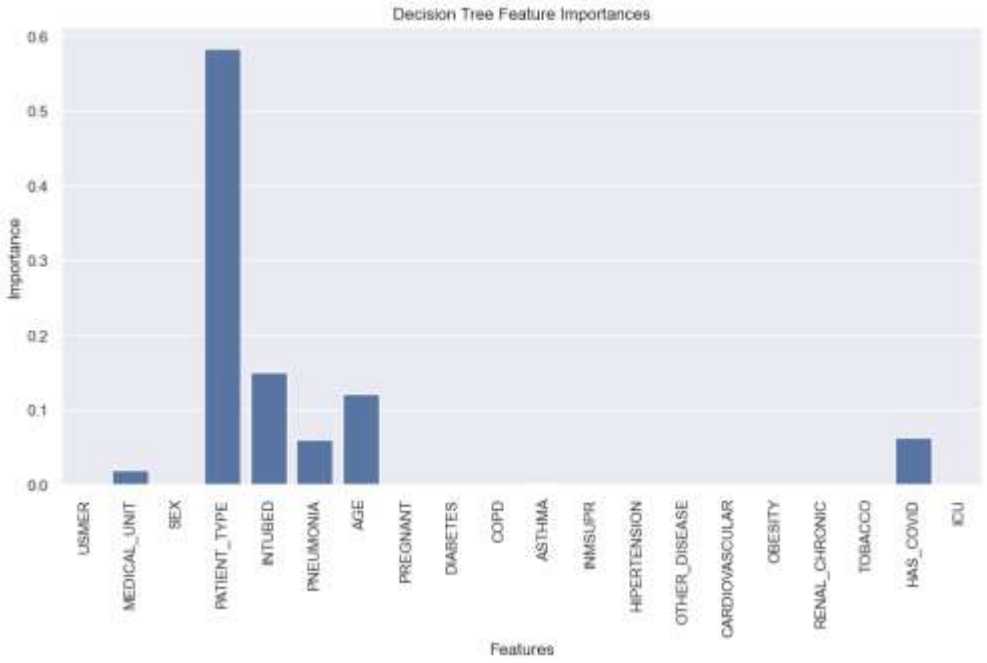
درخت تصمیم

درخت تصمیم یک الگوریتم دسته‌بندی قدرتمند و انعطاف‌پذیر است که داده‌ها را بر اساس مجموعه‌ای از قوانین و شرایط به طبقه‌های مختلف اختصاص می‌دهد. این الگوریتم یک ساختار درختی ایجاد می‌کند که هر گره داخلی آن یک ویژگی ورودی را نشان می‌دهد. هر شاخه نشان‌دهنده یک قاعده است و هر برگه یک نتیجه دسته‌بندی است. درخت تصمیم می‌تواند برای هر دو وظیفه طبقه‌بندی و رگرسیون استفاده شود و می‌تواند داده‌های گسسته و پیوسته را مدیریت کند. این الگوریتم می‌تواند به طور خودکار ویژگی‌های مهم را انتخاب کند و به راحتی توسط انسان قابل تفسیر است. با این حال، درخت تصمیم ممکن است مستعد بیش‌برازش باشد و ممکن است در برابر تغییرات کوچک در داده‌های آموزشی حساس باشد. با استفاده از ماژول درخت تصمیم در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل درخت تصمیم معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

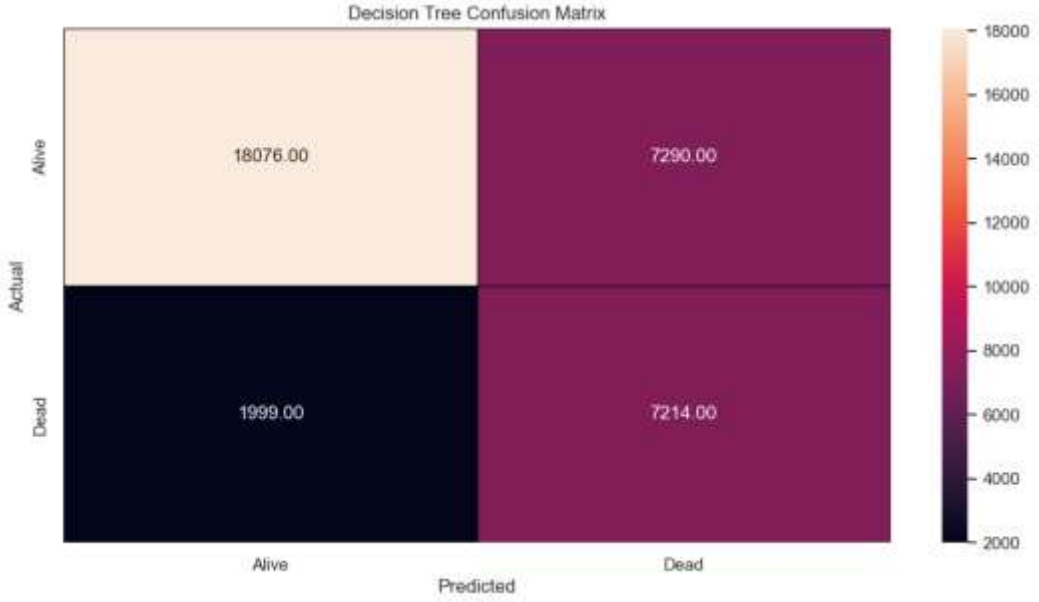
Decision Tree Classification Report:
Training Time: 0.305s | Prediction Time: 0.006s

	precision	recall	f1-score	support
False	0.90	0.71	0.80	25366
True	0.50	0.78	0.61	9213
accuracy			0.73	34579
macro avg	0.70	0.75	0.70	34579
weighted avg	0.79	0.73	0.75	34579

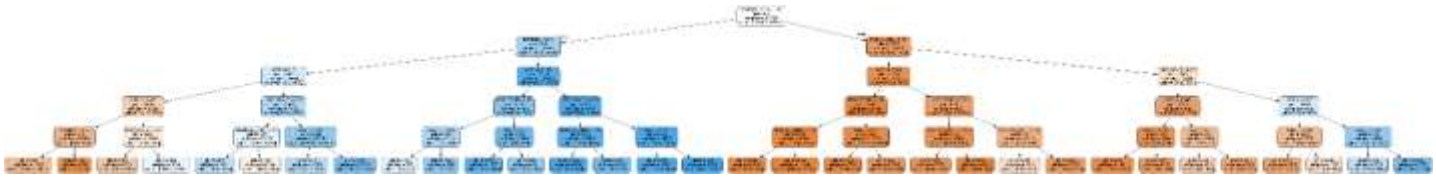
در این نمودار، میزان اهمیت هر ویژگی را داخل درخت تصمیم آموزش‌دیده‌شده می‌توانیم مشاهده کنیم.



همچنین ماتریس درهم‌ریختگی این مدل نیز به شکل زیر می‌باشد.



همچنین تصویری از درخت رسم شده در شکل زیر مشاهده می‌شود که برای بررسی بیشتر در کنار فایل تمرین قرار داده خواهد شد.

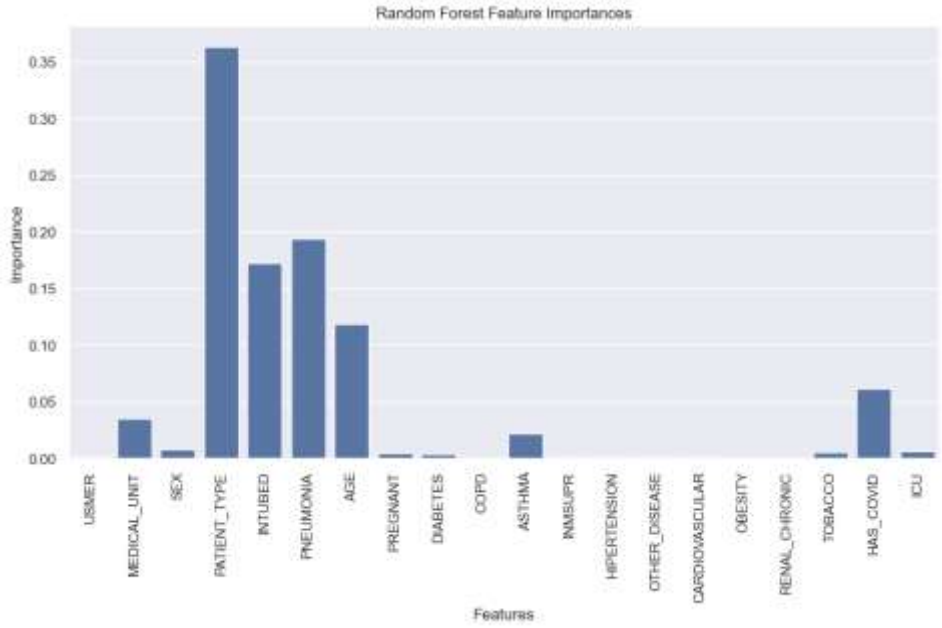


جنگل تصادفی

جنگل تصادفی یکی از الگوریتم‌های یادگیری ماشین است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. این الگوریتم یک مجموعه‌ای از درخت‌های تصمیم را آموزش می‌دهد و پیش‌بینی‌های خود را با میانگین‌گیری یا رای‌گیری ترکیب می‌کند. این روش به دلیل کاهش خطای عمومی در مقایسه با یک درخت تصمیم منفرد، انعطاف‌پذیری و توانایی مدیریت داده‌های گسسته و پیوسته، بسیار قدرتمند است. جنگل تصادفی می‌تواند با تعداد زیادی از ویژگی‌ها و داده‌ها سازگار شود و در برابر نویز و داده‌های از دست رفته مقاوم است. این الگوریتم در طیف وسیعی از برنامه‌های کاربردی از جمله تشخیص تصاویر، پیش‌بینی مالی و توصیه‌های شخصی‌سازی‌شده استفاده می‌شود. با استفاده از ماژول جنگل تصادفی در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل جنگل تصادفی معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Random Forest Classification Report:
Training Time: 1.884s | Prediction Time: 0.032s

              precision    recall  f1-score   support

       False       0.92      0.68      0.78     25366
        True       0.48      0.83      0.61      9213

    accuracy                           0.72     34579
   macro avg       0.70      0.75      0.69     34579
weighted avg       0.80      0.72      0.73     34579
```

در این نمودار، میزان اهمیت هر ویژگی را در الگوریتم جنگل تصادفی می‌توانیم مشاهده کنیم.



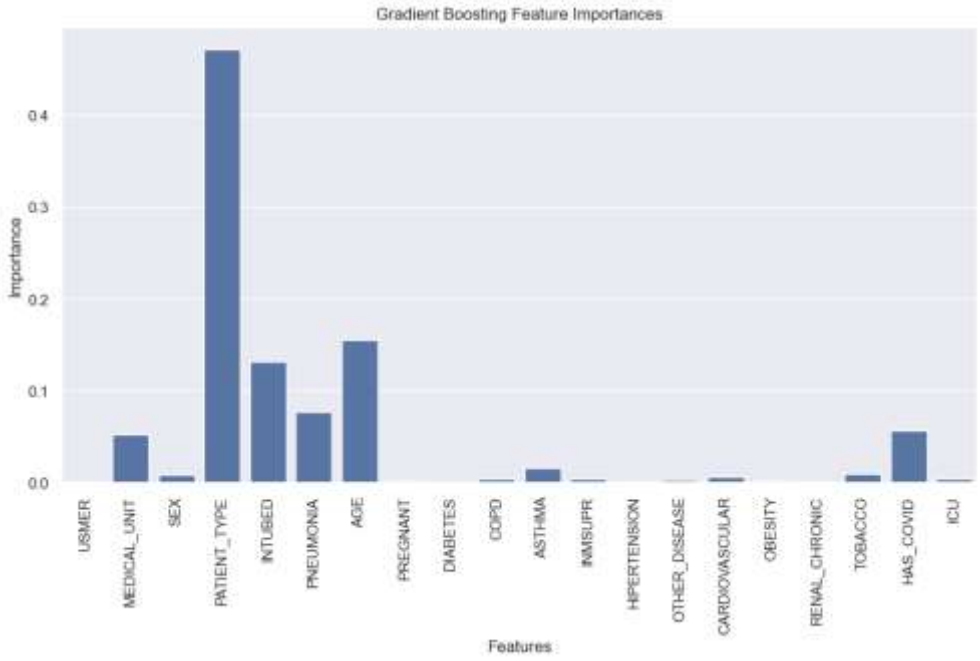
همچنین ماتریس درهم‌ریختگی این مدل نیز به شکل زیر می‌باشد.

تقویت گرادیان

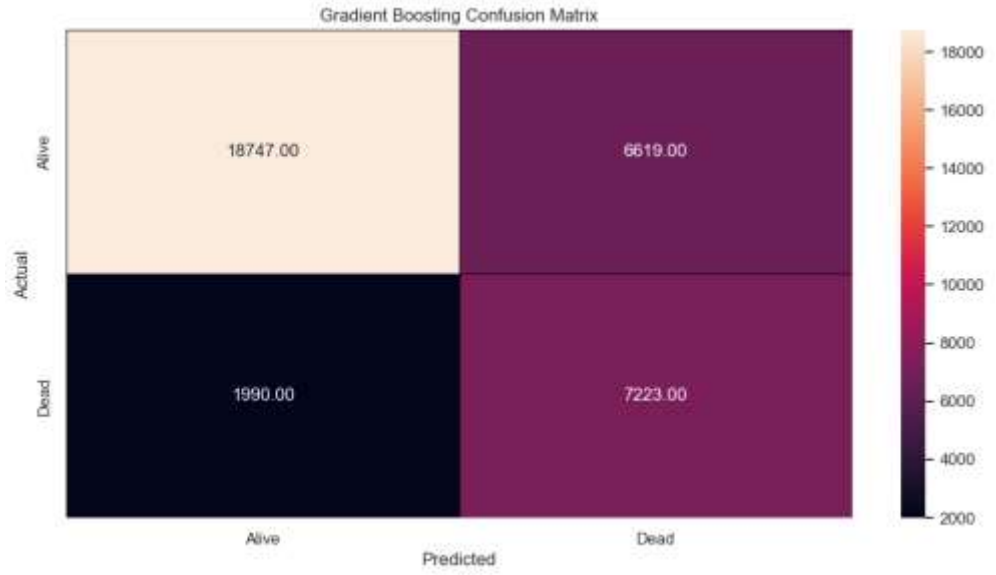
تقویت گرادیان یک الگوریتم یادگیری ماشین قدرتمند و انعطاف‌پذیر است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. این روش شامل آموزش یک سری مدل‌های ضعیف است که هر کدام بر اساس خطای مدل قبلی آموزش داده می‌شوند. تقویت گرادیان از گرادیان کاهشی برای به حداقل رساندن خطا استفاده می‌کند و می‌تواند با روش‌های منظم‌سازی برای جلوگیری از بیش‌برازش ترکیب شود. با استفاده از ماژول دسته‌بند تقویت گرادیان در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل تقویت گرادیان معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
Gradient Boosting Classification Report:
Training Time: 14.968s | Prediction Time: 0.049s

              precision    recall  f1-score   support

       False       0.90      0.74      0.81     25366
        True       0.52      0.78      0.63      9213

    accuracy                           0.75     34579
   macro avg       0.71      0.76      0.72     34579
weighted avg       0.80      0.75      0.76     34579
```

در این نمودار، میزان اهمیت هر ویژگی را در الگوریتم تقویت گرادیان می‌توانیم مشاهده کنیم.



همچنین ماتریس درهم‌ریختگی این مدل نیز به شکل زیر می‌باشد.

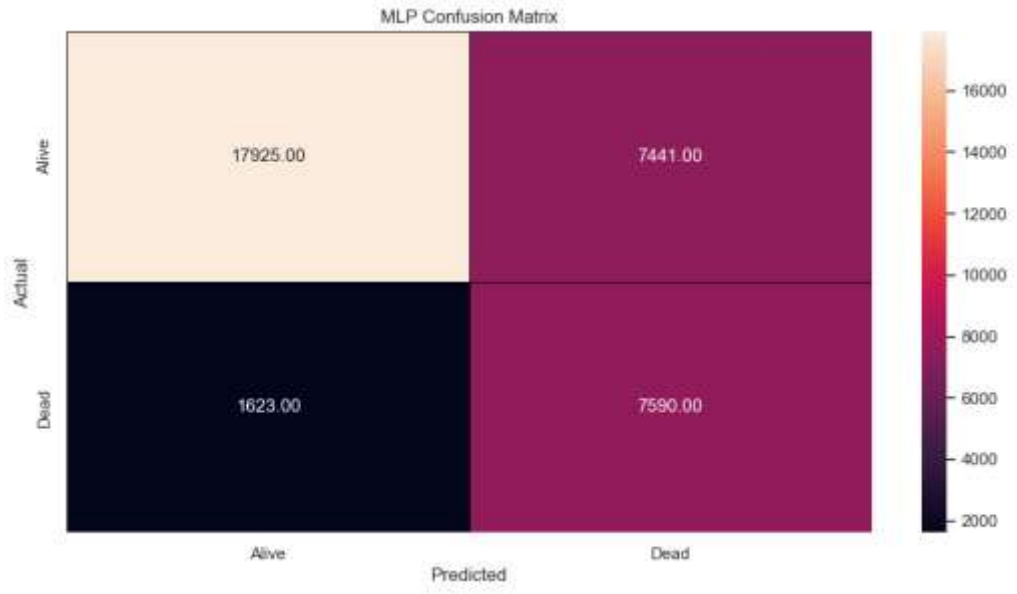


شبکه پرسپترون چند لایه

شبکه پرسپترون چند لایه یک مدل یادگیری عمیق است که از شبکه‌های عصبی مصنوعی برای دسته‌بندی استفاده می‌کند. شبکه پرسپترون چند لایه از چندین لایه نورون‌های متصل تشکیل شده است که داده‌های ورودی را پردازش می‌کنند و از طریق تابع فعال‌سازی مانند ReLU عبور می‌دهند. این شبکه‌ها می‌توانند الگوها و روابط پیچیده‌ای را بین ورودی‌ها و خروجی‌ها یاد بگیرند و برای دسته‌بندی دودویی یا چنددسته استفاده شوند. شبکه‌های پرسپترون چند لایه با استفاده از الگوریتم پس‌انتشار خطا آموزش داده می‌شوند که گرادیان خطا را برای به‌روزرسانی وزن‌ها محاسبه می‌کند. با استفاده از ماژول دسته‌بند شبکه پرسپترون چند لایه در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل شبکه پرسپترون چند لایه معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

```
MLP Classification Report:
Training Time: 64.150s | Prediction Time: 0.025s

              precision    recall  f1-score   support

       False       0.92      0.71      0.80     25366
        True       0.50      0.82      0.63      9213

    accuracy                           0.74     34579
   macro avg       0.71      0.77      0.71     34579
weighted avg       0.81      0.74      0.75     34579
```

همچنین ماتریس درهم‌ریختگی این مدل نیز به شکل زیر است.

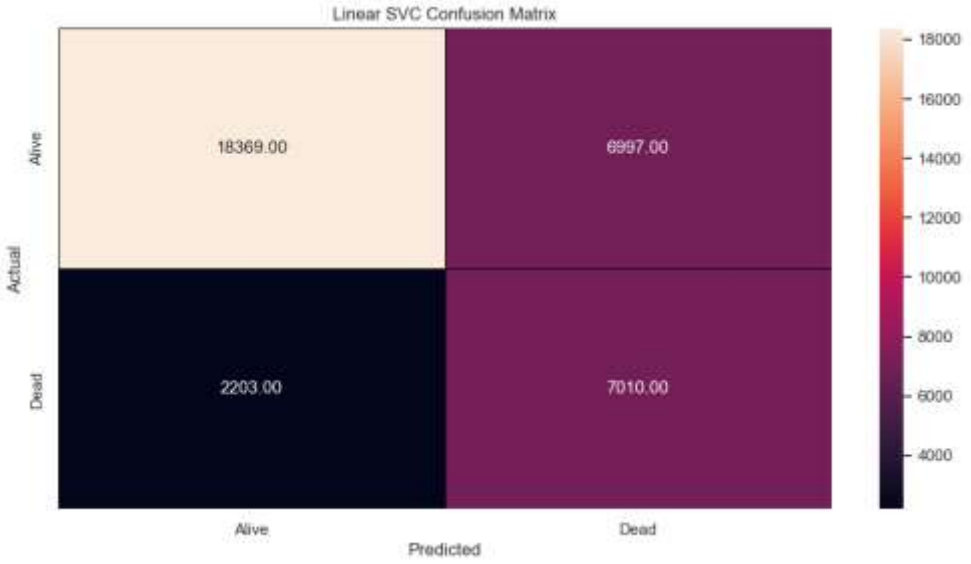


ماشین بردار پشتیبان

Linear SVC یک الگوریتم دسته‌بندی ماشین بردار پشتیبان است که برای داده‌های خطی بهینه شده است. ماشین بردار پشتیبانی به طور معمول از یک تابع هسته برای تبدیل ویژگی‌های غیرخطی استفاده می‌کند، اما Linear SVC این تابع هسته را حذف می‌کند و فرض می‌کند که داده‌ها به طور خطی قابل جداسازی هستند. این ساده‌سازی باعث می‌شود که Linear SVC بسیار سریع‌تر از ماشین بردار پشتیبان استاندارد باشد، به ویژه زمانی که تعداد زیادی نمونه یا ویژگی وجود داشته باشد. Linear SVC برای داده‌های با ابعاد بالا مناسب است. با استفاده از ماژول دسته‌بند Linear SVC در کتابخانه ذکر شده یک مدل را آموزش داده و معیارهای دقت (Precision)، بازیابی یا حساسیت (Recall) و امتیاز F1 را بر روی داده‌های تست ارزیابی می‌کنیم. در مدل Linear SVC معیارهای ذکر شده بر روی داده‌های تست به شرح زیر می‌باشند.

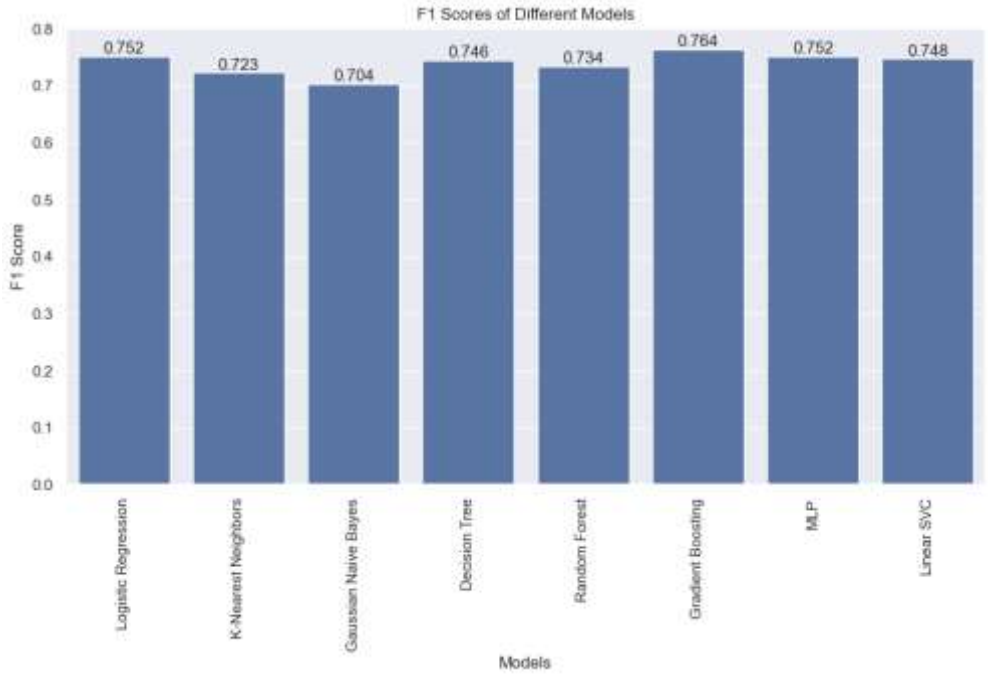
```
Linear SVC Classification Report:
Training Time: 0.536s | Prediction Time: 0.006s

              precision    recall  f1-score   support

       False       0.89      0.72      0.80     25366
        True       0.50      0.76      0.60      9213

    accuracy                           0.73     34579
   macro avg       0.70      0.74      0.70     34579
weighted avg       0.79      0.73      0.75     34579
```

همچنین ماتریس درهم‌ریختگی این مدل نیز به شکل زیر می‌باشد.

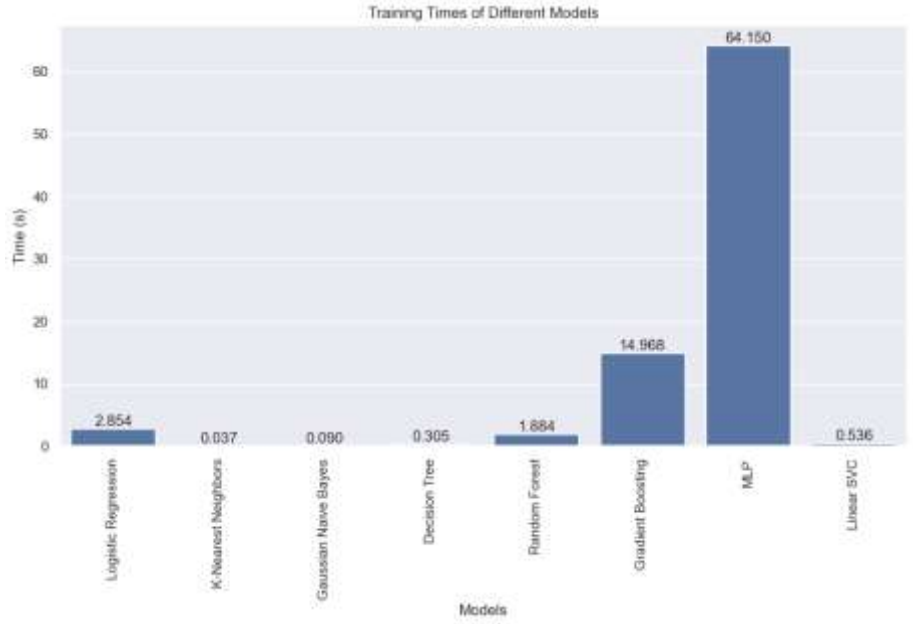


مقایسه مدل‌های یادگیری ماشین

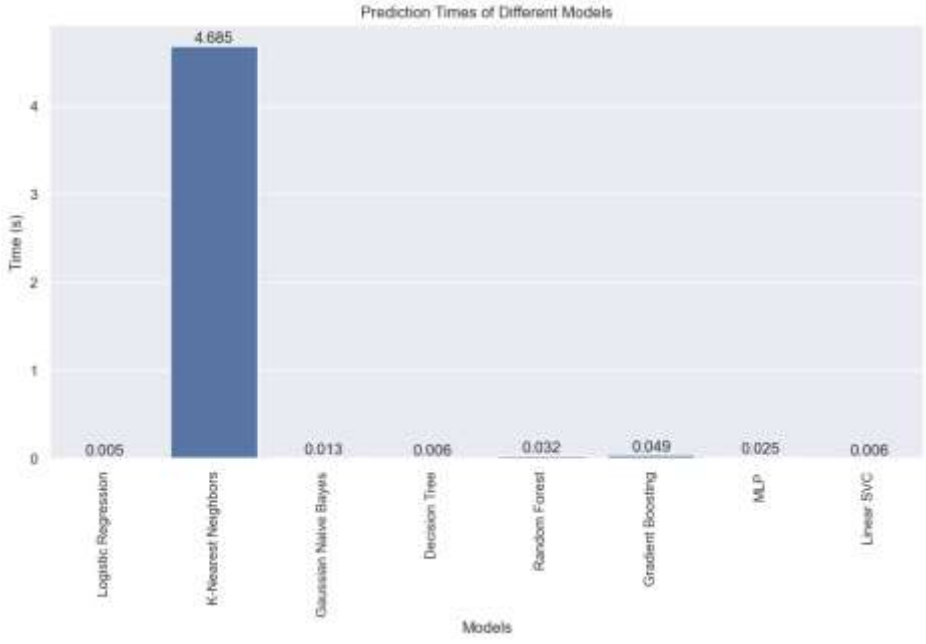
در این بخش با استفاده از معیارهای مشخص شده و همچنین نمودارهای رسم شده به بررسی عملکرد روش‌های مختلف می‌پردازیم. در نمودار زیر، مقایسه‌ای را بین امتیاز F1 کسب شده توسط مدل‌ها بر روی مجموعه ارزیابی انجام داده‌ایم. مشاهده می‌شود که تقویت گرادیان بالاترین امتیاز را کسب کرده است.



همچنین، میان زمان آموزش هر مدل نیز مقایسه‌ای انجام شده است.



در نهایت، برای زمان استنتاج هر مدل نیز یک مقایسه‌ای انجام شده است.



با توجه به نتایج حاصل شده، جهت ساخت سیستمی در این زمینه، استفاده از الگوریتم تقویت گرادیان با توجه به زمان کم آموزش و استنتاج و دقت بالایی که کسب کرده است می‌تواند مناسب‌ترین گزینه باشد.

منابع

مجموعه داده بیماران دارای تومور مغزی

https://www.kaggle.com/datasets/thegoanpanda/brain-tumor-stage-based-recurrence-patterns

مجموعه داده کوید ۱۹

https://www.kaggle.com/datasets/meirnizri/covid19-dataset

مستندات کتابخانه scikit-learn

https://scikit-learn.org/stable/index.html

تعاریف الگوریتم‌های یادگیری ماشین

https://www.wikipedia.org/