چکیده‌ای از طرح تحقیقاتی

عنوان طرح:

کاربرد روش های جنگل تصادفی بقا در بیماران مبتلا به سرطان پستان در پیش بینی اولین متاستاز و مقایسه با تحلیل رگرسیون کاکس

Application of random survival forest in breast cancer patients in prediction first metastasis and comparison with cox regression analysis

کلید واژه‌ها:

رگرسیون کاکس-جنگل تصادفی بقا-سرطان پستان

Key Words:

Cox regression- Random survival forest- Breast cancer

مجری اصلی:

لاین تحقیقاتی مجری اصلی:

مقدمه و ضرورت اجرای طرح (به صورت خلاصه):

سرطان پستان پس از سرطان پوست دومین سرطان شایع در زنان است. هر ساله تعداد زیادی از مبتلایان به سرطان پستان تشخیص داده می شوند. در حدود سه چهارم از موارد بیماری در زنان بالای ۵۰ سال ایجاد می‌شود. نرخ مرگ بر اثر این بیماری در کشورهای توسعه یافته برابر با ۱۲/۶% است. لازم به ذکر است که، حدود یک درصد ازکل سرطان‌های پستان در مردان جوان رخ می‌دهد.علت اصلی سرطان پستان هنوز مشخص نیست. هورمون‌های زنانه وافزایش سن، بخشی ازاین نقش را ایفا می‌کند. سرطان پستان در زنان بالای ۵۰ سال بیماری شایعی است. در بین امریکایی‌ها اگر خانم‌ها تا سنین بالا زندگی کنند حداقل۱ نفر از هر ۸ نفر دچار سرطان پستان می‌شوند. احتمال سرطان پستان در خانم‌های بین ۴۰-۳۰ سال حدود ۱ در ۲۵۰ می‌باشد. احتمال سرطان در بین خانم‌های ۵۰-۴۰ سال ۱ در ۷۰ می‌باشد. سرطان پستان، مانند دیگر سرطان‌ها، با استفاده از مشخصه‌های گوناگون به صورت‌های مختلفی تقسیم‌بندی می‌شود. یکی از این تقسیم‌بندی‌ها با استفاده از مرحلهٔ سرطان انجام می‌شود. مرحلهٔ سرطان بیانگر شدت و گستردگی بیماری است و به ۵ رسته تقسیم می‌شود که آن‌ها را با شماره‌های ۴ - ۰ مشخص می‌کنیم. عدد صفر نشان‌گر کمترین شدت و گستردگی بیماری و عدد ۴ نشان‌گر بیشترین شدت و گستردگی بیماری است. تعیین مرحلهٔ سرطان بر اساس معیارهای زیر انجام می‌گیرد،

۱. اندازهٔ تومور (T)

۲. این‌که آیا تومور در سطح گره‌های لنفاوی در زیربغل پخش شده است یا خیر (N)
۳. آیا تومور در قسمت‌های دیگر بدن گسترش یافته است یا خیر (M)

که به اختصار به آن تقسیم‌بندی TNM می‌گویند

هر چه معیارهای فوق افزایش پیدا کند، مرحلهٔ سرطان بالاتر رفته و بیماری شدیدتر خواهد بود.

در جدول زیر تشکیل مراحل مختلف بر اساس تقسیم‌بندی TNM آمده است.

	T	N	M
Stage 0	Tis	N0	M0
Stage I	T1	N0	M0
Stage II			
(Stage IIA)	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
(Stage IIB)	T2	N1	M0
	T3	N0	M0
Stage III			
(Stage IIIA)	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1, N2	M0
(Stage IIIB)	T4	Any N	M0
	Any T	N3	M0
Stage IV	Any T	Any N	M1

مرحلهٔ سرطان بر حسب تقسیم‌بندی TNM

ممکن است سرطان از تک سلولی که از نظر زنتیکی غیرطبیعی است، آغاز گردد. با تکثیر این سلول تومور حاصل می‌شود و منبع خونی برای تغذیه و تداوم رشدش ایجاد می کند. گاهی اوقات سلول ها از توده اولیه جدا شده و در فرآیند "متاستاز" به قسمت های دیگر بدن منتشر می شوند.

زمانیکه سلول های سرطانی به سایر مکان های بدن ازطریق رگ های خونی و یا رگ های لنفاوی حرکت می کند سرطان پستان گسترش می یابد. بیشترین مکان گسترش، غدد لنفاوی منطقه ایی است .این غدد لنفاوی ممکن است غدد زیر بغلی ،گردنی و یا بالای ترقوه ایی باشند. شایعترین محل متاستاز دور شامل استخوان ها، شش ها و کبد می باشد در حالی که متاستاز سرطان پستان به مغز

کمتر رخ می دهد. سرطان می تواند مجددا در پوست همان پستان که درگیرشده (اگر درحین درمان برداشته نشده باشد) و دیگر بافت های قفسه سینه و یا سایر قسمت های بدن بطور موضع عود کند.

بیشتر مواقع قبل از رخ دادن متاستاز، سرطان پستان تشخیص و درمان می شود. بر اساس جدیدترین داده ها از موسسه ملی سرطان ۶۱% سرطان های پستان، قبل از متاستاز تشخیص داده می شوند و ۳۱% از آنها بعد از گسترش به غدد لنفاوی نزدیک یا خارج سینه و ۶ %از آنها بعد از متاستاز دور به غدد لنفاوی نزدیک تشخیص داده می شوند.

مرحله متاستاز در سرطان به معنی گسترش تومور اولیه در محل دور دست بدن است بطوریکه با رسیدن این بیماری به مرحله متاستاز، سلول های سرطانی از محل تومور اولیه به نقاط دوردست بدن پخش شده و قسمت های دیگر بدن را درگیر می کند.

اخیرا جنگل های تصادفی بقا(RSF) برای تحلیل داده های بقا استفاده می شودکه این یک روش درختی کلی (جمعی)(ensemble tree method) برای تحلیل داده های بقا سانسور از راست می باشد. ساختن مجموعه ای از ساختارهای درختی می تواند به طور معناداری عملکرد یادگیری (learning performance)را بهبود بخشد(Ishwaran & Kogalur, 2007). مین (Minn)و همکاران (۲۰۰۷)علامت بیان ژنی متاستازهای ریه که در میان متاستازهای سرطان سینه واقع شده اند را مورد تحقیق و وارسی قرار داده اند. آن ها از RSF جهت تعیین اثر فاکتورهای خطر متاستاز بیماران استفاده کردند. ایشواران وکوگالور [1]برای تحلیل داده هایی از یک کارآزمایی بالینی تصادفی شده از بیماران از روش RSF استفاده کردند و مقدار لگ رتبه ای، بقای رخدادها و قواعد جداسازی لگ رتبه ای تقریبی را روی این مجموعه داده هابا هم مقایسه کردند[۱].

هدف این مطالعه مقایسه عملکرد روش های تحلیل رگرسیون کاکس (CRA) وجنگل های تصادفی بقا (RSF) با داده های واقعی مربوط به سرطان پستان می باشد.

اهداف و فرضیات

الف. هدف کلی:

تحلیل داده های مربوط به بیماران سـرطان پسـتان با اسـتفاده از روش های جنگل تصادفی بقا و تحلیل رگرسیون کاکس ومقایسه آن ها

ب. اهداف جزیی (اختصاصی):

۱. بکارگیری روش جنگل تصـادفی بقا(RSF) جهت کشـف متغیر های پیشگوی مهم وتاثیر گذار در بروز متاستاز در بیماران مبتلا به سرطان پستان

[1] - Ishwaran and Kogalur

۲. بکارگیری روش کاکس(CRA) جهت کشف متغیر های پیشگوی مهم وتاثیر گذار در بروز متاستاز در بیماران مبتلا به سرطان پستان

ج. هدف فرعی:

مقایسه عملکرد روشهای (RSF) و(CRA) با استفاده از نمونه موجود و شبیه سازی آن در حجم نمونه های مختلف

مقایسه روشهای (RSF) و(CRA) در پیشگویی فاکتورهای مهم وتاثیر گذار

د. هدف کاربردی:

کشف اثرات مهم و تاثیر گذار در بروز اولین متاستاز در بیماران مبتلا به سرطان پستان

هـ سؤالات پژوهشی و فرضیات:

آیا دو روش رگرسیون کاکس (CRA) و جنگل تصادفی بقا(RSF) جهت رده بندی وجداسازی متغیرهای مهم وموثر در بروز اولین متاستاز دقت یکسانی دارند؟

آیاروشهای (RSF) و(CRA) با شبیه سازی در حجم نمونه های مختلف عملکرد یکسانی دارند؟

استفاده از روش جنگل تصادفی بقا (RSF) برای داده های بقا با تعداد متغیرهای پیشگوی زیاد، یک الگوی رده بندی و پیشگویی مناسب ارائه می دهد.

روش جنگل تصادفی بقا (RSF)عملکرد بهتری در کشف متغیرهای مهم وتاثیر گذار دارد.

روش اجرا و شیوه‌های تجزیه و تحلیل یافته‌ها (به صورت خلاصه):

جامعه آماری در مطالعه حاضر که یک مطالعه طولی گذشته نگر است افراد مورد مطالعه، بیماران مبتلا به سرطان پستان می باشند که طی سالهای ۷۷ الی۸۹ (بازه زمان تشخیص: ۱۳۷۷/۸/۱۴ تا ۸۹/۶/۳۰) به درمانگاه آنکولوژی بیمارستان سیدالشهداء اصفهان مراجعه کرده اند و تحت درمان قرار گرفته اند. زمان پایان مطالعه ۲۴ اردیبهشت ۱۳۹۰ بوده است. بنابراین تمام بیماران مراجعه کننده

و مبتلا به سرطان پستان که تعداد آنها ۱۰۸۵ نفر می باشد، وارد مطالعه شده اند. در بین این تعداد از بیماران، ۸۸ نفر در هنگام ورود به مطالعه دارای متاستاز می باشند،معیار ورود به مطالعه نداشتن متاستاز می باشد در بین این تعداد از بیماران، ۸۸ نفر در هنگام ورود به مطالعه دارای متاستاز می باشند، لذا این تعداد از بیماران از مطالعه حاضر خارج شدند و حجم نمونه به ۹۹۷ کاهش یافت. ومعیار خروج وقوع متاستاز برای بیمار می باشد. اطلاعات بیماران در فرم های مخصوصی که توسط مرکز تحقیقات سرطان پستان طراحی گردیده وارد و در پرونده های پزشکی بیماران ثبت شده است. اطلاعات استخراج شده برای این مطالعه دربرگیرنده اطلاعات فردی واطلاعات وابسته به تومور بیمار می باشد. اطلاعات فردی بیمار شامل اطلاعاتی از قبیل سن بیمار و سابقه فامیلی سرطان پستان می باشد که از طریق مصاحبه از بیمار پرسیده می شود. اطلاعات وابسته به تومور عبارتند از سایز تومور، تعداد غده های لنفاوی درگیر زیر بغل، تعداد غده های لنفاوی برداشته شده از زیر بغل، وضعیت گیرنده هورمونی استروژن، گیرنده هورمونی پروژسترون، ژن P53، ژن HER2، کاتپسین- د وKi67 که در گزارش پاتولوژی موجود بوده است. سایر اطلاعات از قبیل بافت درگیر در متاستاز و زمان بروز متاستاز ها توسط پزشک گزارش شده است.

در این مطالعه بیماران تا وقوع متاستاز های متوالی پیگیری شدند. هدف از این مطالعه تعیین ریسک فاکتورهای مهم در بروز اولین متاستاز براساس دو روش رگرسیون کاکس و جنگل تصادفی بقا و هم چنین دسته بندی عوامل خطر ساز در وقوع اولین متاستاز در بیماران براساس این دو روش و مقایسه آن ها و شبیه سازی در حجم نمونه های مختلف می باشد.

مبانی نظری از طریق جستجو در پایگاه های اطلاعاتی اعم از کتاب ها، پایان نامه ها و مقالات موجود در این زمینه جمع آوری خواهد شد. با استفاده از چک لیست مرتبط اطلاعات لازم در مورد متغیرهای زمینه ای، مستقل و زمان اولین متاستاز بیماران از پرونده ها استخراج می شود. تحلیل داده ها بر اساس مدل مورد نظر و اجرای محاسبات با استفاده از نرم افزار R وبا استفاده از پکیج های چون ipred و rsf برای اجرای مدل انجام خواهد شد.

جنگل های تصادفی یک نوع مدرن از روش های درخت –پایه هستند که شامل انبوهی از درخت های کلاس بندی و رگرسیونی اند. مهم ترین ویژگی جنگل های تصادفی عملکرد بالای آن ها در اندازه گیری اهمیت متغیرها برای مشخص کردن اینکه هر متغیر چه نقشی در پیش بینی پاسخ دارد می باشد. جنگل های تصادفی برای گروه بندی افراد با ناهمگنی بکار می روند. اما انگیزه اصلی ما در این مطالعه بکار بردن جنگل های تصادفی برای کشف پیشگوهای مهمی می باشد که می توانند در شانس ابتلا به بیماری موثر باشند.

کاربرد روش های جنگل تصادفی بقا در بیماران مبتلا به سرطان پستان در پیش بینی اولین متاستاز و مقایسه با تحلیل رگرسیون کاکس

ب) عنوان طرح به انگلیسی:

Application of random survival forest in breast cancer patients in prediction first metastasis and comparison with cox regression analysis

نوع طرح:

بنیادی بنیادی- کاربردی کاربردی ■

۱-مقدمه

سرطان پستان پس از سرطان پوست دومین سرطان شایع در زنان است. هر ساله تعداد زیادی از مبتلایان به سرطان پستان تشخیص داده می شوند. در حدود سه چهارم از موارد بیماری در زنان بالای ۵۰ سال ایجاد می‌شود. نرخ مرگ بر اثر این بیماری در کشورهای توسعه یافته برابر با ۱۲/۶% است. لازم به ذکر است که، حدود یک درصد ازکل سرطان‌های پستان در مردان جوان رخ می‌دهد.علت اصلی سرطان پستان هنوز مشخص نیست. هورمون‌های زنانه وافزایش سن، بخشی ازاین نقش را ایفا می‌کند. سرطان پستان در زنان بالای ۵۰ سال بیماری شایعی است. در بین امریکایی‌ها اگر خانم‌ها تا سنین بالا زندگی کنند حداقل ۱ نفر از هر ۸ نفر دچار سرطان پستان می‌شوند. احتمال سرطان پستان در خانم‌های بین ۴۰-۳۰ سال حدود ۱ در ۲۵۰ می‌باشد. احتمال سرطان در بین خانم‌های ۵۰-۴۰ سال ۱ در ۷۰ می‌باشد. سرطان پستان، مانند دیگر سرطان‌ها، با استفاده از مشخصه‌های گوناگون به صورت‌های مختلفی تقسیم‌بندی می‌شود. یکی از این تقسیم‌بندی‌ها با استفاده از مرحلهٔ سرطان انجام می‌شود. مرحلهٔ سرطان بیانگر شدت و گستردگی بیماری است و به ۵ رسته تقسیم می‌شود که آن‌ها را با شماره‌های ۴ - ۰ مشخص می‌کنیم. عدد صفر نشان‌گر کمترین شدت و گستردگی بیماری و عدد ۴ نشان‌گر بیشترین شدت و گستردگی بیماری است. تعیین مرحلهٔ سرطان بر اساس معیارهای زیر انجام می‌گیرد،

۴. اندازهٔ تومور (T)
۵. این‌که آیا تومور در سطح گره‌های لنفاوی در زیربغل پخش شده است یا خیر (N)
۶. آیا تومور در قسمت‌های دیگر بدن گسترش یافته است یا خیر (M)

که به اختصار به آن تقسیم‌بندی TNM می‌گویند

هر چه معیارهای فوق افزایش پیدا کند، مرحلهٔ سرطان بالاتر رفته و بیماری شدیدتر خواهد بود.

در جدول زیر تشکیل مراحل مختلف بر اساس تقسیم‌بندی TNM آمده است.

	T	N	M
Stage 0	Tis	N0	M0
Stage I	T1	N0	M0
Stage II			
(Stage IIA)	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
(Stage IIB)	T2	N1	M0
	T3	N0	M0
Stage III			
(Stage IIIA)	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1, N2	M0
(Stage IIIB)	T4	Any N	M0
	Any T	N3	M0
Stage IV	Any T	Any N	M1

مرحلهٔ سرطان بر حسب تقسیم‌بندی TNM

ممکن است سرطان از تک سلولی که از نظر زنتیکی غیرطبیعی است، آغاز گردد. با تکثیر این سلول تومور حاصل می‌شود و منبع خونی برای تغذیه و تداوم رشدش ایجاد می کند. گاهی اوقات سلول ها از توده اولیه جدا شده و در فرآیند "متاستاز" به قسمت های دیگر بدن منتشر می شوند.

زمانیکه سلول های سرطانی به سایر مکان های بدن ازطریق رگ های خونی و یا رگ های لنفاوی حرکت می کند سرطان پستان گسترش می یابد. بیشترین مکان گسترش، غدد لنفاوی منطقه ایی است .این غدد لنفاوی ممکن است غدد زیر بغلی ،گردنی و یا بالای ترقوه ایی باشند. شایعترین محل متاستاز دور شامل استخوان ها، شش ها و کبد می باشد. متاستاز سرطان پستان به مغز کمتر رخ می دهد. سرطان می تواند مجددا در پوست همان پستان که درگیرشده (اگر درحین درمان برداشته نشده باشد) و دیگر بافت های قفسه سینه ویاسایرقسمتهای بدن بطورموضع عودکند.

بیشتر مواقع قبل از رخ دادن متاستاز، سرطان پستان تشخیص و درمان می شود. بر اساس جدیدترین داده ها از موسسه ملی سرطان ۶۱% سرطان های پستان، قبل از متاستاز تشخیص داده می شوند و ۳۱% از آنها بعد از گسترش به غدد لنفاوی نزدیک یا خارج سینه و ۶ %از آنها بعد از متاستاز دور به غدد لنفاوی نزدیک تشخیص داده می شوند.

مرحله متاستاز در سرطان به معنی گسترش تومور اولیه در محل دور دست بدن است بطوریکه با رسیدن این بیماری به مرحله متاستاز، سلول های سرطانی از محل تومور اولیه به نقاط دوردست بدن پخش شده و قسمت های دیگر بدن را درگیر می کند.

تحلیل بقا مجموعه ای از روش های آماری برای تحلیل داده هایی است که متغیر پاسخ آن ها زمان لازم، تارخداد یک پیشامد می باشد. منظور از زمان در این تعریف می تواند سال، ماه، هفته و یا زمان شروع یک مطالعه تا زمان رخداد پیشامد مورد نظر، ویا سن فرد در زمان رخداد پیشامد مورد نظر می باشد و منظور از پیشامد می تواند مرگ، بروز بیماری، عود بیماری پس از فروکش کردن نشانه های آن(relapse from remission)، بهبودی (recovery) باشد و یا بطور کلی هر تجربه ی تعریف شده ای باشد که با آن مواجه می شوند. در تحلیل بقا معمولا به متغیر زمان، عنوان زمان بقا (survival time) داده می شود [۱۹].

عبارت داده‌های بقا[1] برای توصیف داده‌هایی است که زمان تا وقوع پیشامد [2] خاصی را اندازه‌گیری می‌کنند. زمان تا وقوع پیشامد، متغیری با مقادیر مثبت و دارای توزیع پیوسته است.

مدل‌های رگرسیونی به‌کار گرفته شده برای داده‌های بقا، به‌طور سنتی برپایهٔ مدل مخاطرات متناسب کاکس بوده است که به ابزاری پرکاربرد برای تحلیل رگرسیونی داده‌های سانسور شده تبدیل شده است[۲].

برخی اوقات، محقق با داده‌هایی روبرو می‌شود که در آن متغیر برآمد زمان بقا است. مهمترین ویژگی این نوع داده‌ها، سانسور شدگی است. در داده‌های سانسور شده، زمان دقیق رخداد پیشامد مشخص نیست. هنگامی که با این نوع داده‌ها سروکار داریم نمی‌توانیم از مدل‌های رگرسیونی خطی و لجستیک استفاده کنیم، زیرا در هیچ‌یک از این دو مدل خاصیت سانسور شدگی داده‌ها مدنظر قرار نمی‌گیرد. روش رگرسیونی‌ای که برای مدل‌سازی داده‌های بقا بکار گرفته می‌شود کمی متفاوت از رگرسیون کلاسیک است. مطالعاتی که تا کنون در زمینهٔ مدل‌سازی داده‌های بقا بکار گرفته شده‌اند عمدتاً به روش‌های پارامتری و نیمه- پارامتری پرداخته‌اند. روش‌های پارامتری، که در آن رابطهٔ بین متغیرهای کمکی و زمان پیشامد به‌طور کامل مشخص می‌شود، در رشته‌های مهندسی بسیار موفق عمل می‌کنند، زیرا مکانیسم تحت مطالعه کاملاً شناخته شده است. در کاربردهای زیستی و پزشکی، مدل‌های پارامتری نسبت به مدل‌های نیمه-پارامتری از محبوبیت کمتری برخوردارند، زیرا، برخلاف رشته‌های مهندسی، در این حالت اطلاعات پایه‌ای واضحی در اختیار نداریم. در مدل‌های رگرسیونی نیمه- پارامتری، ساختار وابستهٔ متغیرهای کمکی روی زمان شکست مشخص است در حالی که توزیع مخاطرهٔ پایه مشخص نیست. رستهٔ اصلی مدل‌های نیمه- پارامتری در تحلیل بقا، مدل مخاطرات متناسب کاکس است.

1 -Survival Data
2 -Time to Event

تعریف واژه‌ها:

سرطان پستان چیست؟

سرطان پستان، رشد مهار نشده‌ی سلول‌های غیر طبیعی است که در نواحی مختلف پستان ایجاد می شود. این اتفاق ممکن است در بافت‌های مختلف مانند مجاری‌ای که شیر را انتقال می‌دهند، در بافت تولید کننده‌ی شیر و در بافت غیر غددی رخ دهد(۲

متاستاز سرطان پستان:

. مرحله متاستاز در سرطان به معنی گسترش تومور اولیه در محل دور دست بدن است بطوریکه با رسیدن این بیماری به مرحله متاستاز، سلول های سرطانی از محل تومور اولیه به نقاط دوردست بدن پخش شده و قسمت های دیگر بدن را درگیر می کند.

رگرسیون کاکس:

معمولاًدر مطالعات بقا ،حادثه مد نظر "مرگ" می باشد .در این مطالعات عده ای را برای مدتی تحت مراقبت وپیگیری قرار می دهند تا تعدادی "مرگ"بدلیل خاصی که مد نظر وبررسی مطالعه است مشاهده گردد .

بدیهی است که عوامل و فاکتورهای دیگری چون متغیرهای جمعیتی نظیر سن وجنس بیمار، متغیرهای روحی و روانی، عوامل جسمانی و فیزیکی خصوصاً وضعیت قلبی ، متغیرهای محیطی وسابقه سیگاری بودن وعادات غذایی بر حادثه مشخص فوق تاثیر دارند. از طرف

دیگر بعضی از بیماری ها، علل مستقیم تری دارند که می توانند زمان بقا را قویاً تحت تأثیر خود قرار دهند. مثلاً در مطالعه بیماران سرطانی، سن بیمار، اندازه غده، نحوه درمان و... تأثیر مستقیم بر روند وقوع مرگ دارند.

بنابراین برای ضرورت مدل سازی داده های بقا می توان دو دلیل عمده زیر را بیان کرد:

۱) مدل سازی، ترکیب متغیرهای موثر بر تابع خطر را مشخص می نماید.

۲) برآورد تابع خطر برای هر فرد مشخصی و در هر لحظه زمانی را می توان با مدل سازی مناسب بدست آورد .

مدل اساسی که برای داده های بقا به کار برده می شود"رگرسیون کاکس" می باشد که به نام مدل خطرات متاسب نیز شناخته شده است.

به منظور سهولت هرچه بیشتر، مدل خطرات متناسب کاکس را برای حالتی شروع می کنیم که خطرات غیر متناسب و یا متغیرهای غیر توجیهی مرتبط با زمان در مدل وجود نداشته باشد. بنابراین اگر**k** متغیرمستقل $x_1,x_2,x_3,\ldots,x_k$ داشته باشیم، تابع خطر در زمان **t** به صورت زیر تعریف می شود.

$$h_i(t)=h_0(t).\exp\{\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}\}$$

براساس مدل فوق، مخاطره پایه $h_0(t)$ تابع مخاطره برای فردی است که مقدار متغیر توضیحی اش برابر صفر است. چون فرض نشده که این تابع مخاطره پایه فرمی پارامتریک دارد، مدل کاکس را یک مدل نیمه پارامتریک برای تابع مخاطره گویند. تابع بقای متناسب با این مدل بصورت زیر خواهد بود:

$$S(t;x)=\exp[-\exp(x'\beta)\int_0^t \lambda_0(u)\,du$$

روش های درخت- پایه[1]

روشهای درخت- پایه روش های آماری ناپارامتری (مدل آزاد) برای اجرای آنالیز کلاس بندی و آنالیز رگرسیونی با استفاده از الگوریتم افراز های بازگشتی می باشند [۳و۴]. این روش ها در انتخاب یک مجموعه از متغیرهای پیشگو، که به بهترین نحو فنوتیپ نهایی بیماری رابیان کنند، بسیار کارا هستند. روش های درخت-پایه وقتی که متغیرهای پیشگو بطور غیر خطی در ارتباط با بیماری هستند نیز مفیدند چون هیچ قیدی را در مورد فرم رابطه بین متغیرهای پیشگو و پاسخ فرض نمی کنند. این روش ها با اغلب ناهمگنی های ژنتیکی[2] سازگارند (ناهمگنی ژنتیکی به این معناست که راه های ممکن متعددی برای ایجاد یک بیماری وجود دارد که هرکدام شامل زیر مجموعه

1 - Tree-based
2 -genetic heterogeneity

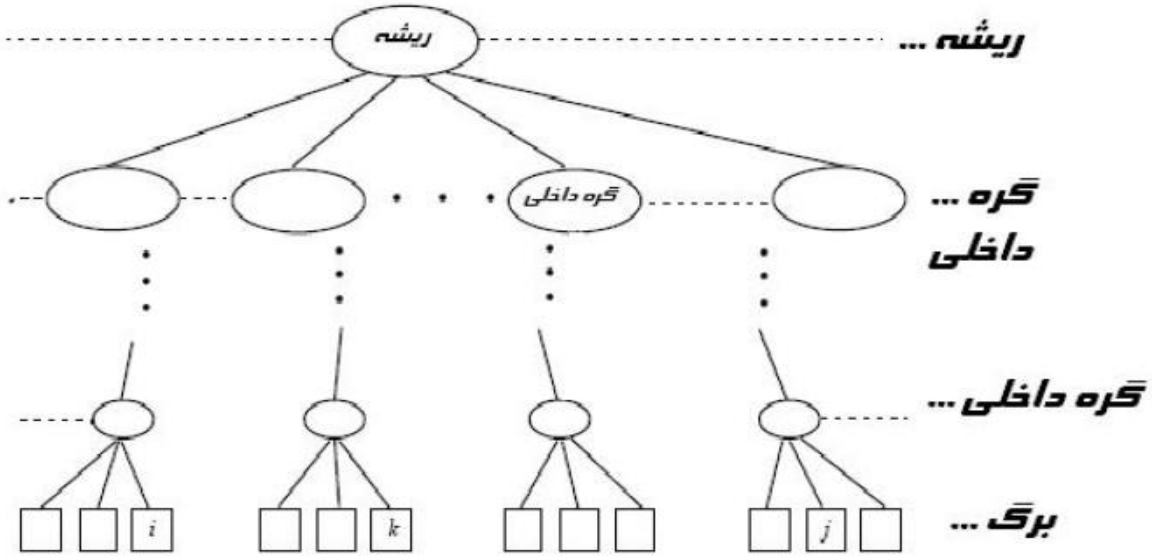
های مختلفی از ژن هاست). بدین صورت که به طور اتوماتیک مدل های جداگانه ای به زیر مجموعه هایی از داده ها، که با افراز زود هنگام در درخت مشخص می شوند، برازنده می شوند. سادگی مدل و قابل تفسیر بودن روش های درخت-پایه، انعطاف پذیری در بکارگیری تعداد زیاد متغیر های پیشگو و حجم نمونه محدود و توانایی شان در مد نظر قرار دادن ناهمگنی ژنتیکی منجر به افزایش کاربرد آنها در مطالعات همبستگی شده است.

درخت تصمیم[1]

درخت تصمیم یکی از روش های ناپارامتری رده بندی کردن می باشد. این روش بابکارگیری تکنیک های بسیارساده، یک الگوی رده بندی را برای مشاهدات موجود معرفی می نماید. الگوی معرفی شده توسط این روش، از ساختاری بسیار ساده و قابل درک برای تصمیم گیری برخوردار می باشد. با اینکه این روش از تکنیک های ساده ای استفاده می نماید ولی در زمینه تشخیص و پیشگویی می تواند به خوبی روش های پیچیده ای نظیر شبکه های عصبی مصنوعی عمل نماید. درخت تصمیم یک روش ساده و توانمند برای طبقه بندی یک مجموعه به رده های متمایز و همگن می باشد که یک گراف غیر چرخشی شبیه درخت دارد که این درخت توسط مجموعه ای از سوالات نشان داده می شود و معمولا هر سوال با توجه به یک متغیر مطرح می شود.

یک گراف درخت تصمیم از سه جز اصلی ریشه[2] ، گره داخلی[3] و گره خارجی[4] (برگ) تشکیل شده است و روند بدین گونه است که ابتدا یک متغیر کمکی به عنوان ریشه انتخاب می گردد و با توجه به یک سری از سوالات و ویژگی ها به چندین گره داخلی تقسیم می شود. به تقسیمات انجام شده یک زیر درخت می گویند. به عنوان نمونه مشاهدات براساس گروه های سنی مختلف به چند گروه تقسیم می شوند ویا اینکه براساس جنسیت به دو گروه تقسیم می شوند. هر گره داخلی نیز مانند ریشه به گره های دیگری تقسیم می شود تا در نهایت به هر گره یک رده از متغیر پاسخ منتسب می شود که این گره ها برگ نامیده می شوند. ساختار کلی از یک درخت تصمیم در نمودار زیر نشان داده شده است[۵].

1 - Decision Tree
2 - Root
3 - Internal Node
4- External Node(Leaf)



نمایش ساختارکاملی از یک درخت تصمیم

درخت تصمیم در بسیاری از زمینه ها به کار برده می شود که یکی از کاربردهای آن در زمینه تشخیص و پیشگویی در علوم پزشکی است. ولی می توان به جرات گفت که مهم ترین جایگاه و کاربرد درخت تصمیم در داده کاوی[1] است یعنی آن جایی که به قدری با مشاهدات و متغیرهای گوناگون سروکار داریم که نمی توان به سادگی از روش های کلاسیک و موجود استفاده کرد. برای تحلیل این نوع از اطلاعات و داده ها به روش هایی نیازمندیم که بتوان خیلی سریع و با انجام محاسبات کمتر به نتایج قابل قبول دست یافت. یکی از راه هایی که این امکان را فراهم می نماید استفاده از درخت تصمیم است. هر گره تصمیم درخت را کاندیدی برای هرس در نظر می گیریم. هرس یک گره تصمیم شامل حذف زیردرخت منتج به آن گره و تبدیل آن گره به یک گره برگ و انتساب رایج ترین دسته بندی برای آن گره به عنوان دسته بندی گره موردنظر می باشد. گره ها فقط زمانی حذف می شوند که درخت بدست آمده از عمل هرس، بدتر از درخت اولیه روی مجموعهء تایید عمل نکند. همیشه گره ای برای هرس انتخاب می شود که حذف آن بیشترین افزایش را در دقت درخت تصمیم روی نمونه داشته باشد. گره ها مکرراً هرس شده و این عمل تا جایی ادامه می یابد که هرس کردن بیشتر دقت درخت را کاهش دهد.

روش های معمول آماری به سه دلیل عمده نمی توانند در تشخیص و پیشگویی های پزشکی مورد استفاده قرار گیرند.

۱. در زمینه اطلاعات پزشکی معمولا با مقادیر بسیاری از متغیرهای کمکی سروکار داریم که این حجم زیاد متغیرها استفاده از روش های معمول را بامشکل و پیچیدگی هایی روبرو می سازد. به عنوان مثال اثرات متقابل در این داده ها به قدری پیچیده است که نمی توان با روش های معمول آماری آن ها را بررسی نمود.

[2]-Data Mining

۲. روش های معمول آماری نیازمند وجود بعضی از پیش فرض ها نظیر نرمال بودن توزیع متغیرها و واریانس های همگن می باشند که در داده های پزشکی معمولا چنین فرضیاتی برقرار نخواهد بود.

۳. نتایج بدست آمده از روش های معمول آماری به سادگی قابل استفاده نمی باشند.

درخت تصمیم یکی از شیوه های مناسب رده بندی است که با وجود مشکلات ذکر شده می تواند به خوبی عمل نماید [۱۷].

معمولا برای معرفی یک درخت تصمیم مشاهدات موجود به طور تصادفی به دو دسته نمونه یادگیری یا فراگیری[1] و نمونه آزمون[2] تقسیم می گردد. به طور کلی برای شناخت و معرفی یک درخت تصمیم مناسب و کارا باید سه مرحله راطی نماییم.

۱. باید شرایط ومعیارهایی را برای ساخت و رشد یک درخت تصمیم معرفی کنیم.

۲. باید سعی کنیم درختی با اندازه کوچک تر ولی با همان دقت یا بیشتر را معرفی نماییم.

۳. باید مناسب بودن درخت معرفی شده را بررسی نماییم.

جنگل های تصادفی[3] (RF)

جنگل های تصادفی یک نوع مدرن از روش های درخت –پابه هستند که شامل انبوهی از درخت های کلاس بندی و رگرسیونی اند[۶]. مهم ترین ویژگی جنگل های تصادفی عملکرد بالای آن ها در اندازه گیری اهمیت متغیرها برای مشخص کردن اینکه هر متغیر چه نقشی در پیش بینی پاسخ دارد می باشد. جنگل های تصادفی برای گروه بندی افراد با ناهمگنی بکار می روند. اما انگیزه اصلی ما در این مطالعه بکار بردن جنگل های تصادفی برای کشف پیشگوهای مهمی می باشد که می توانند در شانس ابتلا به بیماری موثر باشند.

یک RF مجموعه ای از درخت های هرس نشده است که هر درخت با الگوریتم افرازهای بازگشتی بدست می آید.

الگوریتم ساخت یک RF با T درخت از یک مجموعه داده با n مشاهده و p متغیر بدین صورت است:

i. باروش بوت استرپ(یک روش محاسباتی- آماری- کامپیوتری است برای تعیین میزان دقت برآوردگرهای حاصل از داده نمونه است. دراین تکنیک تنها با یک روش خیلی ساده می‌توان تقریباً هر آماره‌ای از توزیع داده‌های نمونه را تخمین زد. این روش از روشهای باز نمونه گیری به حساب می‌آید) یک نمونه تصادفی باجایگذاری به حجم n از مشاهدات انتخاب می شود.هر نمونه بوت استرپ از داده ها یک **OOB**نامیده می شود.

[1]- Training or Learning sample
[2]- Test sample
[3] - Random Forests

ii. برای نمونه بوت استرپ انتخاب شده یک درخت کلاس بندی با استفاده از الگوریتم افرازهای بازگشتی، رشد می کند. در هرگره افراز براساس یک نمونه تصادفی m تایی از p متغیر پیشگو انجام می شود. گره با استفاده از متغیر کاندیدی که تفاوت بقا بین گره ها را ماکزیمم می کند، انشعاب پیدا می کند.

iii. الگوریتم افرازهای بازگشتی آن قدر ادامه می یابد تا درخت به بزرگترین اندازه خود(یعنی برای هر مشاهده یک گره نهایی)، برسد بدون آن که درخت هرس شود.

iv. مراحل (i) تا (iii) T بار تکرار می شوند تا یک RF ساخته شود (۵). انتخاب های رایج برای T، ۱۰۰۰ درخت و برای m ، $\sqrt{P}$ و log(p) هستند[۷].

یک RF آن قدر بزرگ است که تفسیر آن کار بسیار دشواری است، لذا نیازمند خلاصه کردن اطلاعات آن با استفاده از شاخص های کمی هستیم. یکی از این شاخص ها اهمیت متغیر[۱] **(VI)** است. VI شاخصی برای رتبه بندی متغیرها بر حسب اهمیت آن ها در اثر گذاری روی پاسخ است. معروف ترین شاخص های VI ، شاخص اهمیت جینی [۲] و شاخص اهمیت جایگشتی[۳] می باشد.

در **RF** تصادفی سازی در دو بخش تعریف می شود در بخش اول به صورت تصادفی یک نمونه بوت استرپ از داده ها بدست می آید و از آن جهت رشد درخت استفاده می شود. در بخش بعدی در هر گره درخت یک زیر مجموعه ای از متغیرها که به طور تصادفی انتخاب شده اند به عنوان متغیرهای کاندید جهت انشعاب سازی انتخاب می شوند.

میانگین گیری از درختان در ترکیب با تصادفی سازی در رشد درخت استفاده می شود و هم چنین**RF** را قادر می سازد تا کلاس های قدرتمندی از توابع در حالی که خطای تعمیم کمی بدست می آورد ، تقریب بزند.

بالاخره درخت بقا به یک نقطه اشباع می رسد وقتی که هیچ مورد دیگری نتواند به خاطر این معیار که هر گره باید شامل حداقل $d_0 > 0$ مرگ باشد، شکل گیرد.بیشتر گره ها در یک درخت اشباع شده گره های نهایی نامیده می شود.

فرض کنید $(T_{1,h},....,\delta_{1,h})$ **,.....,** $(T_{n(h),h},....,\delta_{n(h),h})$ زمان های بقا باشند فرد **i**ام در زمان $T_{i,h}$ از راست سانسور است اگر $\delta_{i,h}$ = ۰ باشد.

اگر $t_{1,h} < t_{2,h} < ... < t_{n(h),h}$ ،N(h)زمان گسسته بقا باشند و $Y_{l,h}$, $d_{1,h}$ به ترتیب تعداد مرگ ها و تعداد افراد در زمان $t_{1,h}$ باشند برآورد CHFبرای hبرآوردگر نلسون آلن می باشد.

---

1 - Variable importance
2 -Gini importance index
3- Permutation importance index

$$\hat{H}_h(t) = \sum_{t_{l,h} \le t} \frac{d_{l,h}}{Y_{l,h}}$$

در مطالعه که توسط نوری و همکاران انجام شد، از روش آماری ناپارامتری و نوین جنگل های تصادفی برای تعیین فاکتورهای مهم واثر گذار ژنتیکی بر روی بیماران آنکیلوزان اسپوندیلیت استفاده شد. تحلیل های فوق متعاقبا به کمک رگرسیون لجستیک نیز اجرا شد و نتایج آن با جنگل های تصادفی مقایسه گردیدکه هر کدام از این روش های فاکتورهای متفاوتی را به عنوان مهم ترین فاکتور در رابطه با بیماری معرفی کردند[۱۸].

شاخص اهمیت جینی:

در طی ساخت درخت های RF برای تعیین اینکه گره براساس کدام متغیر افراز شود، از شاخص ناخالصی جینی استفاده می شود . اهمیت متغیر X_i در یک درخت مجموع کاهش در شاخص ناخالصی جینی روی تمام گره هایی است که براساس X_i افراز شده اند. میانگین اندازه اهمیت متغیر X_i روی تمام درخت های جنگل ، اندازه شاخص اهمیت جینی است[۸].

شاخص اهمیت جایگشتی:

الگوریتم RF از تمام مشاهدات نمونه برای ساخت درخت استفاده نمی کند بلکه یک نمونه تصادفی با جایگذاری (به حجم معمولا برابر) از مشاهدات انتخاب می شود. به مشاهدات انتخاب شده نمونه آموزشی[1](LS) و به بقیه آن ها نمونه خارج کیسه[2](OOB) گفته می شود. درخت ها با مشاهدات LS ساخته می شوند واز OOB برای اندازه گیری ناخالصی درخت استفاده می شود. در هر درخت ابتدا اندازه ناخالصی روی مشاهدات OOB محاسبه می شود. سپس مقادیر متغیر X_i مشاهداتOOB بطور تصادفی جابه جا می شوند و اندازه ناخالصی درخت روی مقادیر جابه جا شده محاسبه می شوند. اندازه اهمیت متغیر X_i در هر درخت ، اختلاف بین این دو اندازه ناخالصی است و میانگین این مقادیر شاخص اهمیت جایگشتی است. انگیزه این روش این است که اگر X_i متغیر مهمی باشد جابجا شدن مقادیر آن بطور تصادفی منجر به افزایش ناخالصی درخت می شود در حالی که اگر متغیر تاثیر گذاری نباشد، تغییری در ناخالصی ایجاد نمی شود.

بیان مساله و ضرورت اجراح تحقیق:

1 - Learning sample
2 - Out-of-bag_

تحلیل بقا شامل مجموعه ای از روش های آماری برای تحلیل داده هایی می باشد که درآن متغیر برآمد مورد نظر زمان تا وقوع رخداد می باشد. یکی از مشهور ترین تحلیل های بقا رگرسیون کاکس می باشدکه یک روش نیمه پارامتری و یک روش برای تشخیص اثر چندین متغیر روی زمانی که یک رویداد مشخص اتفاق می افتد، می باشد.

. در حدود سه چهارم از موارد بیماری در زنان بالای ۵۰ سال ایجاد می‌شود. نرخ مرگ بر اثر این بیماری در کشورهای توسعه یافته برابر با ۱۲/۶٪ است[۹, ۱۰]. لازم به ذکر است که، حدود یک درصد ازکل سرطان‌های پستان در مردان جوان رخ می‌دهد[۱۱]. علت اصلی سرطان پستان هنوز مشخص نیست. هورمون‌های زنانه وافزایش سن، بخشی ازاین نقش را ایفا می‌کند[۱۲]. سرطان پستان در زنان بالای ۵۰ سال بیماری شایعی است. در بین امریکایی‌ها اگر خانم‌ها تا سنین بالا زندگی کنند حداقل ۱ نفر از هر ۸ نفر دچار سرطان پستان می‌شوند. احتمال سرطان پستان در خانم‌های بین ۴۰-۳۰ سال حدود ۱ در ۲۵۰ می‌باشد. احتمال سرطان در بین خانم‌های ۵۰-۴۰ سال ۱ در ۷۰ می‌باشد[۱۳]. سرطان پستان پس از سرطان پوست دومین سرطان شایع در زنان است. هر ساله تعداد زیادی از مبتلایان به سرطان پستان تشخیص داده می شوند[۱۴و۱۵]

اخیرا جنگل های تصادفی بقا (RSF) برای تحلیل داده های بقا استفاده می شودکه این یک روش درختی یک مرتبه(کلی) ویا جمعی برای تحلیل داده های بقا سانسور از راست می باشد. ساختن مجموعه ای از ساختارهای درختی می تواند به طور معناداری عملکرد یادگیری را بهبود بخشد. مین وهمکاران علامت بیان ژنی متاستازهای ریه را که در میان متاستازهای سرطان سینه واقع شده اند را مورد تحقیق و وارسی قرار داده اند. آن ها از RSF جهت تعیین اثر فاکتورهای خطر متاستاز بیماران استفاده کردند. ایشواران وکوگالور [1]برای تحلیل داده هایی از یک کارآزمایی بالینی تصادفی شده از بیماران از روش RSF استفاده کردند [۱۶].

هدف این مطالعه مقایسه عملکرد روش های تحلیل رگرسیون کاکس (CRA) وجنگل های تصادفی بقا (RSF) با یک سری داده های واقعی مربوط به سرطان پستان می باشد.

با توجه به مروری که به مطالعات گذشته گردید روش جنگل تصادفی بقا یک روش موثر در تعیین فاکتورهای مهم وتاثیر گذار در هنگام داشتن تعداد متغیرهای زیاد می باشد. ازآن جایی که ما در این مطالعه متغیرها وفاکتورهای زیادی را برای بیماران مبتلا به سرطان پستان اندازه گیری کردیم می توان این روش را با روش رگرسیون کاکس که یک روش شناخته شده می باشد مقایسه کنیم

[1] - Ishwaran and Kogalur

اهداف و فرضیات

الف. هدف کلی:

تحلیل داده های مربوط به بیماران سـرطان پسـتان با اسـتفاده از روش های جنگل تصادفی بقا و تحلیل رگرسیون کاکس ومقایسه آن ها

ب. اهداف جزیی (اختصاصی):

۱) کشف یک شکل تابعی از روابط بین متغیرهای کمکی و زمان تا بروز اولین متاستاز

۲) بکارگیری روش جنگل تصادفی بقا(RSF) جهت کشف متغیر های پیشگوی مهم وتاثیر گذار

۳) بکارگیری روش کاکس(CRA) جهت کشف متغیر های پیشگوی مهم وتاثیر گذار

ج. هدف فرعی

مقایسه روشهای (RSF) و(CRA) در پیشگویی فاکتورهای مهم وتاثیر گذار

د. هدف کاربردی:

کشف اثرات مهم و تاثیر گذار در بروز اولین متاستاز در بیماران مبتلا به سرطان پستان

هـ سؤالات پژوهشی و فرضیات:

سوال تحقیق:

آیا دو روش رگرسیون کاکس (CRA) و جنگل تصادفی بقا(RSF) جهت رده بندی وجداسازی متغیرهای مهم وموثر در بروز اولین متاستاز دقت یکسانی دارند؟(که دراینجا معیار مقایسه مقدار آکائیک (**AIC**)می باشد)

آیا روش جنگل تصادفی بقا (RSF)عملکرد بهتری در کشف متغیرهای مهم وتاثیر گذار دارد.

.

فرضیه:

اســـتفاده از روش جنگل تصـــادفی بقا (RSF) برای داده های بقا با تعداد متغیرهای پیشـــگوی زیاد، یک الگوی رده بندی و پیشـــگویی مناسب ارائه می دهد.

روش جنگل تصادفی بقا (RSF)عملکرد بهتری در کشف متغیرهای مهم وتاثیر گذار دارد.

فهرست منابع مورد استفاده در این طرح :

۱- Imran Kurt Omurlu**,** Mevlut Ture**.** The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer**,** Expert Systems with Applications 36 (2009) 8582–8588

۲-Lee, E.T. and J.W. Wang, *Statistical methods for survival data analysis*. 2003: Wiley-Interscience.

3- Hastie T TRFJ, The elements of statistical learning : data mining, inference and prediction, In Springer series in statistics New York , Springer 2001; xvi: p. 533.

۴--Liaw, A. and Wiener, M. (2002). Classification and regression by random Forest. *R news* 2/3 18–22.

۵- L.Breiman, J.Friedman,R.Olshen andC.Stone.(1984)Classificationandregression trees,Chapman and Hall.

۶- Braiman L, Random forests, Machine Learn 2001; 45: p. 5-32.

۷- Genuer R PJMTC, Random Forest: some methodological insights 2008.

۸- Breiman L , Classification and regression trees CA, Wadsworth International Groups 1984

9- Alireza, S., et al., Comparison of breast cancer survival in two populations: Ardabil, Iran and British Columbia, Canada. BMC Cancer. **9**.

10- Nagel, G., et al., The impact of comorbidity on the survival of postmenopausal women with breast cancer. Journal of cancer research and clinical oncology, 2004. **130**(11): p. 6.۶۷۰-۶۴

11- Parvin, Y., et al., Breast cancer treatment and ethnicity in British Columbia, Canada. BMC Cancer. **10**.

12- Pukkala E., Kulmala I., Hovi S.L., Hemminki E., Keskimäki I., Lipworth L., et al. Causes of death among Finnish women with cosmetic breast implants, 1971-2001. Annals of plastic surgery. 2003;51(4):339.

13- Törner, A., Proportional hazards and additive regression analysis of survival for severe breast cancer. Stockholm University, 2004.

۱۴-Fisch, T., et al., Variation in survival after diagnosis of breast cancer in Switzerland. Annals of Oncology, 2005. **16**(12): p. 1882.

15- Rajaeefard, A.R., et al., Survival Models in Breast Cancer Patients. IRCMJ, 2009. **11**(3): p. 295-300

۱۶- Brody J.G., Rudel R.A., Michels K.B., Moysich K.B., Bernstein L., Attfield K.R., et al., Environmental pollutants, diet, physical activity, body size, and breast cancer. Cancer. 2007;109(S12):2627-34.

۱۷- ساکی مالحی،امل (۱۳۸۸).مدل درختی بقاوکاربرد آن در تحلیل زیرگروه های همگن از بیماران مبتلا به سرطان کولورکتال،پایان نامه کارشناسی ارشد،دانشگاه تربیت مدرس.

۱۸- نوری، سحر؛ نوری جلیانی،کرامت؛ آنلیز جنگل های تصادفی:یک روش غربال گری در مطالعات با بعد بالا و کاربرد آن در یک مطالعه همبستگی ژنتیکی جمعیت-پایه؛ مجله دانشگاه علوم پزشکی خراسان شمالی؛(۱۳۹۰)

۱۹- حاجی زاده، ابراهیم؛ اصغری، محمد؛ روش های و تحلیل های آماری با نگاه به روش تحقیق

روش اجرای طرح:

جامعه آماری در مطالعه حاضر یک مطالعه طولی گذشته نگر است. افراد مورد مطالعه، بیماران مبتلا به سرطان پستان می باشند که طی سالهای ۷۷ الی۸۹ (بازه زمان تشخیص: ۱۳۷۷/۸/۱۴ تا ۸۹/۶/۳۰) به درمانگاه آنکولوژی بیمارستان سیدالشهداء اصفهان مراجعه کرده اند و تحت درمان قرار گرفته اند. زمان پایان مطالعه ۲۴ اردیبهشت ۱۳۹۰ بوده است. بنابراین تمام بیماران مراجعه کننده و مبتلا به سرطان پستان که تعداد آنها ۱۰۸۵ نفر می باشد، وارد مطالعه شده اند که معیار ورود به مطالعه نداشتن متاستاز می باشد. در بین این تعداد از بیماران، ۸۸ نفر در هنگام ورود به مطالعه دارای متاستاز می باشند، لذا این تعداد از بیماران از مطالعه حاضر خارج شدند و حجم نمونه به ۹۹۷ کاهش یافت و معیار خروج وقوع متاستاز برای بیمار می باشد. اطلاعات بیماران در فرم های مخصوصی که توسط مرکز تحقیقات سرطان پستان طراحی گردیده وارد و در پرونده های پزشکی بیماران ثبت شده است. در این مطالعه بیماران تا وقوع متاستاز های متوالی پیگیری شدند. هدف از این مطالعه تعیین ریسک فاکتورهای مهم در بروز اولین متاستاز براساس دو روش رگرسیون کاکس و جنگل تصادفی بقا و همچنین دسته بندی عوامل خطر ساز در وقوع اولین متاستاز در بیماران براساس این دو روش و مقایسه آن ها می باشد.

اطلاعات استخراج شده برای این مطالعه دربرگیرنده اطلاعات فردی واطلاعات وابسته به تومور بیمار می باشد. اطلاعات فردی بیمار شامل اطلاعاتی از قبیل سن بیمار و سابقه فامیلی سرطان پستان می باشد که از طریق مصاحبه از بیمار پرسیده می شود. اطلاعات وابسته به تومور عبارتند از سایز تومور، تعداد غده های لنفاوی درگیر زیر بغل، تعداد غده های لنفاوی برداشته شده از زیر بغل، وضعیت گیرنده هورمونی استروژن، گیرنده هورمونی پروژسترون، ژن P53، ژن HER2، کاتپسین- د وKi67 که در گزارش پاتولوژی موجود بوده است. سایر اطلاعات از قبیل بافت درگیر در متاستاز و زمان بروز متاستاز ها توسط پزشک گزارش شده است.

در این مطالعه بیماران تا وقوع متاستاز های متوالی پیگیری شدند. هدف از این مطالعه تعیین ریسک فاکتورهای مهم در بروز اولین متاستاز براساس دو روش رگرسیون کاکس و جنگل تصادفی بقا و هم چنین دسته بندی عوامل خطر ساز در وقوع اولین متاستاز در بیماران براساس این دو روش و مقایسه آن ها و شبیه سازی در حجم نمونه های مختلف می باشد.

مبانی نظری از طریق جستجو در پایگاه های اطلاعاتی اعم از کتاب ها، پایان نامه ها و مقالات موجود در این زمینه جمع آوری خواهد شد. با استفاده از چک لیست مرتبط اطلاعات لازم در مورد متغیرهای زمینه ای، مستقل و زمان اولین متاستاز بیماران از پرونده ها استخراج می شود. تحلیل داده ها بر اساس مدل مورد نظر و اجرای محاسبات با استفاده از نرم افزار R وبا استفاده از پکیج های چون ipred و rsf برای اجرای مدل انجام خواهد شد.

جنگل های تصادفی یک نوع مدرن از روش های درخت –پایه هستند که شامل انبوهی از درخت های کلاس بندی و رگرسیونی اند. مهم ترین ویژگی جنگل های تصادفی عملکرد بالای آن ها در اندازه گیری اهمیت متغیرها برای مشخص کردن اینکه هر متغیر چه نقشی در پیش بینی پاسخ دارد می باشد. جنگل های تصادفی برای گروه بندی افراد با ناهمگنی بکار می روند. اما انگیزه اصلی ما در این مطالعه بکار بردن جنگل های تصادفی برای کشف پیشگوهای مهمی می باشد که می توانند در شانس ابتلا به بیماری موثر باشند.

با توجه به مروری که به مطالعات گذشته گردید روش جنگل تصادفی بقا یک روش موثر در تعیین فاکتورهای مهم وتاثیر گذار در هنگام داشتن تعداد متغیرهای زیاد می باشد. ازآن جایی که ما در این مطالعه متغیرها وفاکتورهای زیادی را برای بیماران مبتلا به سرطان پستان اندازه گیری کردیم می توان این روش را با روش رگرسیون کاکس که یک روش شناخته شده می باشد مقایسه کنیم

جنبه جدید بودن و نوآوری:

در وضعیت هایی که با پایگاه داده های وسیع مواجه هستیم، تحلیل اکتشافی داده ها و استخراج اهم اطلاعات نهفته در آن ها خود یک مسئله مهم تلقی می شود. از طرفی تا کنون هیچ مقایسه ای بین دو روش جنگل تصادفی بقا و رگرسیون کاکس جهت رده بندی و کشف پیشگوی های مهم انجام نشده است.

ملاحظات اخلاقی:

در این مطالعه سعی بر حفظ اسرار بیماران بوده و تمامی ملاحظات اخلاقی براساس (بیانیه هلسینکی) رعایت شده است.

جدول زمانی مراحل اجرا و پیشرفت کار

ردیف	نوع فعالیت	فرد مسئول	طول مدت به ماه	زمان اجرا (ماه)											
				۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	...
۱	جمع آوری داده ها			*											
۲	ورود اطلاعات به کامپیوتر				*										
۳	برنامه نویسی					*									
۴	گزارش						*	*							
۵															

جمع کل: ۵ ماه

جدول متغیرها:

نام متغیر	نقش متغیر	نوع متغیر		واحد اندازه‌گیری	روش اندازه‌گیری
		کمی	کیفی		
سن هنگام تشخیص	توضیحی			سال	پرونده پزشکی بیمار
زمان اولین متاستاز	پاسخ			سال	پرونده پزشکی بیمار
عود	توضیحی			۰و۱	پرونده پزشکی بیمار
زمان عود	توضیحی			سال وماه	پرونده پزشکی بیمار
تعداد متاستازها	مخدوش گر			۱و۲و۳و۴	پرونده پزشکی بیمار
مکان اولین متازستاز	توضیحی			۰و۱و۲و۳و۴و۵	پرونده پزشکی بیمار
زمان تشخیص	توضیحی			سال وماه	پرونده پزشکی بیمار
اندازه تومور	توضیحی			مقیاسی	پرونده پزشکی بیمار
غده های درگیر	توضیحی			مقیاسی	پرونده پزشکی بیمار

زمان شیمی درمانی	مخدوش گر			سال وماه	پرونده پزشکی بیمار
تعداد شیمی درمانی	مخدوش گر			۱و۲و۳و..	پرونده پزشکی بیمار

۱.