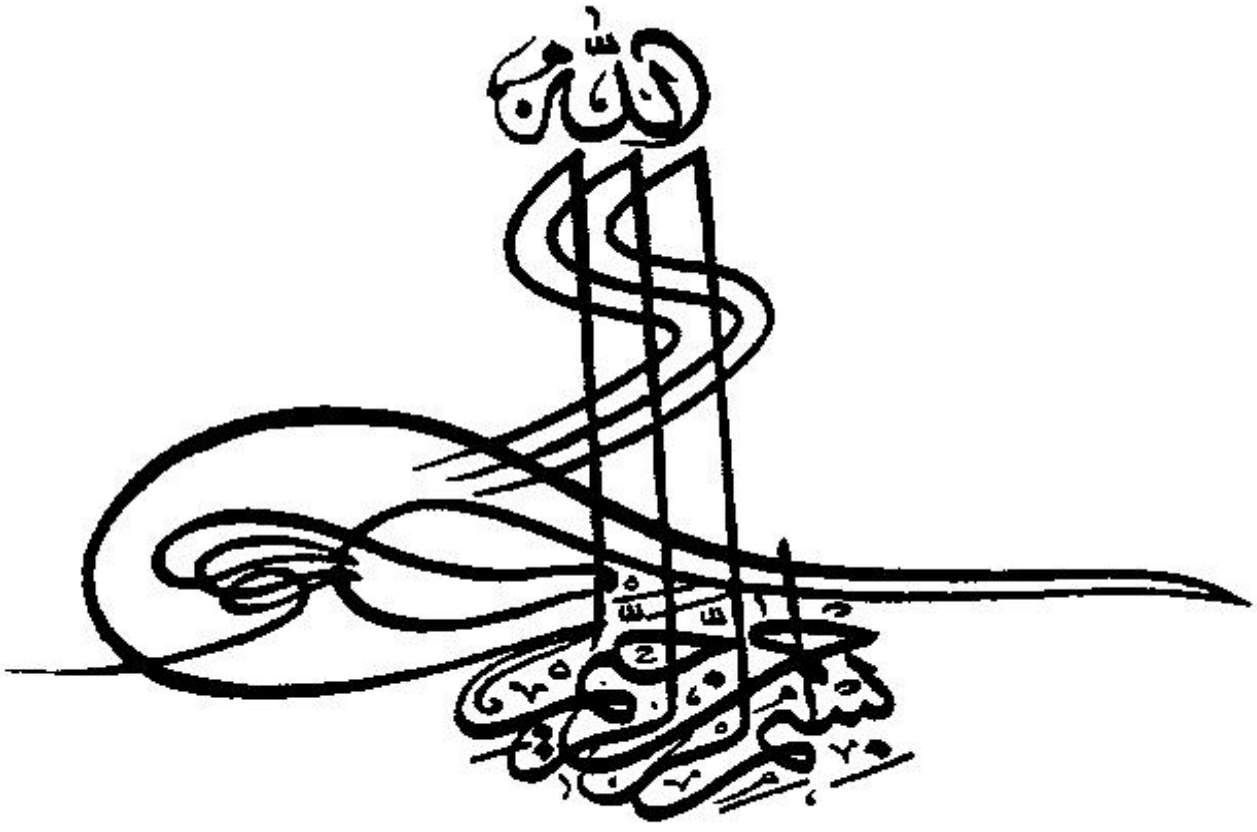

دانشگاه علوم بهزیستی وتوانبخشی

گروه آمار زیستی

پایان نامه کارشناسی ارشد

عنوان

کاربرد شبکه عصبی مصنوعی در پیش بینی پاسخ های دو متغیره آمیخته در بیماری تصلب شرائین

نگارش

مهین عادلی

استاد راهنما

دکتر اکبر بیگلریان

استاد مشاور

دکتر عنایت اله بخشی

استاد مشاور پزشکی

امید علی عادلی

خرداد ۱۳۹۱

شماره ثبت: ۱۰۸-۶۰۰۰

چکیده

مقدمه و اهداف: در مطالعات اپیدمیولوژی و پزشکی، گاهی پژوهش‌گر با مواردی مواجه می‌شود که لازم است دو متغیر پاسخ را به صورت توام (همزمان) از روی تعدادی متغیر کمکی پیش‌بینی نماید. زمانی که متغیر پاسخ آمیخته باشد، با توجه به محدودیت‌ها و برقرار نبودن برخی پیش فرض‌ها، روش‌های کلاسیک آماری برای مدل-بندی و پیش‌بینی کارایی لازم را ندارند. هدف این مطالعه به کارگیری مدل شبکه عصبی مصنوعی برای پیش‌بینی متغیر پاسخ آمیخته در بیماری قلبی است.

روش کار: در پاییز و زمستان ۱۳۹۰، تعداد ۲۷۶ بیمار قلبی که از بیمارستان شهید مدنی خرم آباد ترخیص شده بودند به صورت کوهورت تاریخی مورد مطالعه قرار گرفتند، از این نمونه برای پیش‌بینی توام کلسترول و سطح LDL استفاده شد. داده‌ها به تصادف به دو گروه آموزش(۱۷۵ نفر) و آزمون(۹۱ نفر) تقسیم شدند. برای تحلیل داده-ها از شبکه عصبی مصنوعی با الگوریتم شیب توام مقیاس شده(SCG) و برای تعیین مناسب‌ترین مدل از معیار صحت پیش‌بینی استفاده شد. برای تحلیل از نرم افزار MATLAB نسخه ۷/۱۱ استفاده شد.

نتایج: بالاترین صحت پیش‌بینی برای مدل شبکه عصبی مصنوعی چهار لایه برای متغیر پاسخ آمیخته برابر ۴۱/۷۶ درصد به دست آمد.

نتیجه‌گیری: مدل شبکه عصبی مصنوعی با دو لایه میانی برای پیش‌بینی متغیر پاسخ آمیخته مناسب است.

واژگان کلیدی: شبکه عصبی مصنوعی، پاسخ آمیخته، بیماری قلبی، کلسترول، سطح LDL

تقدیر و تشکر

با سپاس فراوان از دکتر اکبر بیگلریان استاد راهنمای این تحقیق، که از راهنمایی های بی دریغشان در طول انجام این پژوهش و دوران تحصیل بهره برده ام.

با سپاس از دکتر عنایت ا... بخشی و دکتر امید علی عادلی که زحمت مشاوره این تحقیق را عهده دار بودند.

با سپاس از دکتر رهگذر و کریملو به خاطر راهنمایی هایشان در طول دوران تحصیل.

با سپاس از دکتر شیخی ریاست بیمارستان شهید مدنی خرم آباد و کادر این بیمارستان که داده های این پایان نامه را در اختیار این تحقیق قرار دادند.

تقدیم به:

روح بلند پدر بزرگوارم که ریشه آزادگی را در وجودم کاشت

مادرم

همسرم

و تمامی آنان که در مسیر پر پیچ و خم زندگی صفحه ذهنمان را روشن نوشتند

فهرست مطالب

عنوان **صفحه**

فصل اول: کلیات تحقیق ۱

۱-۱.مقدمه و بیان مسئله ۲

۱-۲.اهمیت و ضرورت ۵

۱-۲-۱. آمار مرگ و میر ناشی از بیماریهای قلبی عروقی در ایران و جهان ۶

۱-۳.اهداف پژوهش ۷

۱-۳-۱.هدف کلی..... ۷

۱ -۳-۲.اهداف اختصاصی ۷

۱-۳-۳.اهداف کاربردی ۷

۱-۳-۴.سوالات..... ۸

فصل دوم: مبانی نظری و مروری بر مطالعات گذشته ۹

۲-۱.مقدمه ۱۰

۲-۲.شبکه عصبی مصنوعی ۱۰

۲-۳.شبکه عصبی زیستی ۱۳

۲-۴.طراحی شبکه عصبی مصنوعی ۱۵

۲-۵.ویژگیهای شبکه عصبی مصنوعی ۱۶

۲-۵-۱.قابلیت یادگیری ۱۶

۲-۵-۲.پراکندگی اطلاعات ۱۷
۲-۵-۳.قابلیت تعمیم ۱۸
۲-۵-۴.پردازش موازی ۱۸
۲-۵-۵.مقاوم بودن ۱۸
۲-۶.مدل نرون تک ورودی ۱۸
۲-۷.توابع فعالیت 19
۲-۷-۱.تابع محرک سیگموئید ۲۰
۲-۷-۲.تابع محرک تانژانت هایپربولیک ۲۰
۲-۷-۳.تابع محرک خطی ۲۱
۲-۸.مدل چند ورودی با یک نرون ۲۲
۲-۹.مدل چند ورودی با بیش از یک نرون ۲۳
۲-۱۰.لایه میانی ۲۳
۲-۱۱.انواع شبکه عصبی مصنوعی ۲۴
۲-۱۱-۱.شبکه عصبی مصنوعی پیشخور ۲۴
۲-۱۱-۲.شبکه های عصبی مصنوعی پسخور (بازگشتی) ۲۵
۲-۱۱-۳.شبکه های عصبی مصنوعی رقابتی ۲۵
۲-۱۲.فرایند یادگیری در شبکه عصبی ۲۶
۲-۱۲-۱.تعریف یادگیری ۲۶
۲-۱۲-۲.انواع یادگیری ۲۷

۲- ۱۲-۲-۱.یادگیری با ناظر ۲۷
۲- ۱۲-۲-۱.یادگیری با ناظر ۲۷
۲-۱۲-۲-۳.یادگیری تشدیدی (تقویتی) ۲۸
۲-۱۲-۳.معادلات یادگیری در حالت کلی ۲۸
۲-۱۲-۴.یادگیری شبکه ۲۹
۲-۱۳.شبکه عصبی پرسپترون ۳۰
۲-۱۳-۱.قانون یادگیری پرسپترون ساده (SLPR) ۳۱
۲-۱۳-۲.مشکل پرسپترون ساده ۳۲
۲-۱۳-۳.رفع مشکل پرسپترون ساده ۳۳
۲-۱۳-۴.شبکه عصبی پرسپترون چند لایه (MLP) ۳۴
۲-۱۴.محدودیت شبکه های عصبی مصنوعی ۳۶
۲-۱۵.مدلبندی پاسخهای دومتغیره به وسیله شبکه عصبی مصنوعی ۳۷
۲-۱۶.بررسی متون ۳۸
۲-17.مروری بر مطالعات پزشکی ۴۲
فصل سوم: روش شناسی تحقیق ۴۴
۳-۱.بخش اول : داده های پزشکی ۴۵
۳-۱-۱.مقدمه ۴۵
۳-۱-۲.نوع مطالعه ۴۵
۳-۱-۳.تعریف متغیرها ۴۵

۳-۱-۴.جامعه آماری ، نمونه آماری و روش نمونه گیری ۴۹
۳-۱-۵.روش جمع اوری داده ها ۵۰
۳-۱-۶.متغیر های تحقیق ۵۱
۳-۲. تجزیه و تحلیل داده ها ۵۳
۳-۲-۱.توپولوژی شبکه عصبی مصنوعی برای مدل بندی پاسخ های دو متغیره آمیخته ۵۳
۳-۲-۲.انواع توپولوژی شبکه عصبی مصنوعی برای مدل بندی پاسخ های دو متغیره آمیخته ۵۳
۳-۲-۲-۱.مدل های یک متغیره ۵۳
۳-۲-۲-۲.مدل های دو متغیره ۵۵
۳-۲-۳.الگوریتم یادگیری حداقل میانگین مربعات خطا ۵۶
۳-۲-۴.الگوریتم یادگیری پس انتشار خطا ۵۸
۳-۲-۵.تنظیم پارامتر های لایه میانی ۶۱
۳-۳-۶. پس انتشار حساسیت ها ۶۴
۳-۲-۷.خلاصه الگوریتم BP..... ۶۶
۳-۲-۸.ملاحظاتی در مورد الگوریتم BP ۶۸
۳-۲-۸-۱.نحوه ارائه داده های یادگیری ۶۸
۳-۲-۸-۲. انتخاب مقادیر اولیه ۶۹
۳-۲-۹.الگوریتم BP برای مدل بندی پاسخ های دو متغیره آمیخته ۷۰
فصل چهارم: نتایج و یافته ها ۷۷
۴-۱.مقدمه ۷۸
۴-۲.توصیف داده ها ۷۸

۴-۳.نتایج برازش مدل های شبکه عصبی مصنوعی ۸۲

۴-۳-۱: نتایج برازش مدل های شبکه عصبی مصنوعی با الگوریتم های آموزش متفاوت ۸۲

۴-۳-۲: نتایج برازش مدل های شبکه عصبی مصنوعی با یک لایه میانی ۸۵

۴-۳-۳: نتایج برازش مدل های شبکه عصبی مصنوعی با دو لایه میانی ۸۷

فصل پنجم:بحث، نتیجه گیری و پیشنهادها ۹۶

۵-۱.مقدمه ۹۷

۵-۲.خلاصه یافته ها ۹۷

۵-۳.بحث..... ۹۸

۵-۴.نتیجه گیری ۱۰۱

۵-۵.محدودیت ها ۱۰۱

۵-۵.پیشنهادها ۱۰۱

فهرست منابع ۱۰۲

فهرست جداول

عنوان **صفحه**

جدول۳-۱.فهرست متغیرها ۵۱

جدول ۴-۱. توزیع متغیرهای کیفی مورد بررسی در بیماران قلبی بستری شده در بیمارستان تخصصی قلب شهید مدنی خرم آباد به تفکیک وضعیت کلسترول خون ۷۹

جدول ۴-۲ توصیف متغیرهای کمی مورد بررسی در بیماران قلبی بستری شده در بیمارستان تخصصی قلب شهید مدنی خرم آباد به تفکیک وضعیت کلسترول خون ۸۱

جدول ۴-۳. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE مدل شبکه عصبی مصنوعی سه لایه با توجه به الگوریتم آموزش متفاوت ۸۴

جدول ۴-۴. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به گره های متفاوت در لایه میانی اول ۸۶

جدول ۴-۵. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۷ گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۸۷

جدول ۴-۶. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۱۰گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۸۸

جدول ۴-۷. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۱۳گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۸۹

جدول ۴-۸صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۱۶گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۹۰

جدول ۴-۹. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۱۹ گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۹۱

جدول ۴-۱۰. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۲۲ گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۹۲

جدول ۴-۱۱. صحت پیش بینی متغیرهای پاسخ (به درصد) و MSE با توجه به ۲۵ گره در لایه میانی اول و تعداد گره های متفاوت در لایه میانی دوم ۹۳

جدول ۴-۱۲. خلاصه جدول های ۴-۵ تا ۴-۱۱ به صورت صحت پیش بینی به درصد (تعداد گره لایه میانی اول) ۹۴

فهرست شکل ها

عنوان **صفحه**

شکل ۲-۱.سلول عصبی زیستی ۱۴

شکل ۲-۲. شکل کلی یک شبکه عصبی مصنوعی ۱۵

شکل ۲-۳.نمای کلی یک شبکه عصبی تک ورودی ۱۹

شکل ۲-۴.تابع سیگموئید ۲۰

شکل ۲-۵.تابع تانژانت هایپربولیک ۲۱

شکل ۲-۶.تابع فعالیت خطی ۲۱

شکل ۲-۷.مدل شبکه عصبی چند ورودی با یک نرون ۲۲

شکل ۲-۸ مدل شبکه عصبی چند ورودی با بیش از یک نرون ۲۳

شکل ۲-۹. شبکه عصبی پرسپترون چند لایه ۳۵

فصل اول

کلیات تحقیق

۱–۱: مقدمه و بیان مساله

شناسایی الگو و طبقه بندی یکی از مهمترین کاربردهای روشهای آماری در علوم مختلف است. یکی از اهداف عمده مدلسازی و طبقه بندی در آمار، پیش بینی براساس واقعیت ها، پارامترهای موجود و اطلاعات در دسترس است. دستیابی به این اهداف با استفاده از روشهایی مانند رگرسیون، تحلیل ممیزی، سریهای زمانی، رده بندی و رگرسیون درختی و سایر روشهای آماری امکان پذیر است. در آمار از رگرسیون برای مدلبندی روابط بین متغیرها استفاده می شود. رگرسیون چندمتغیره[1] حالتی است که برای هر مشاهده یک بردار پاسخ چند متغیره وجود دارد. این پاسخها می توانند به صورت کمی، کیفی و یا آمیخته باشند. در حالتی که کلیه متغیرهای پاسخ[2] کمی هستند، روش تحلیل داده ها در آمار، استفاده از مدلهای رگرسیون چند متغیره است در حالتی که همه پاسخها کیفی باشند، روش پیشنهادی استفاده از مدلهای رگرسیون لجستیک چندمتغیره یا مدلهای رگرسیون لجستیک چندمتغیره تعمیم یافته است. اما در حالت سوم که متغیرهای پاسخ، برخی کمی و برخی کیفی هستند، با توجه به محدودیتهایی که برای در نظر گرفتن توزیع مناسب برای بردار پاسخ در این گونه موارد وجود دارد، روشهای کلاسیک در این مورد کارایی لازم را ندارند[۱].

روشهای کلاسیک رگرسیون برای مدلبندی روابط بین متغیرها دارای مفروضاتی مانند: مشخص بودن توزیع متغیرهای پاسخ، خطی بودن رابطه پیشنهادی در برخی مدلهاو یکسان بودن واریانس خطاها هستند. حال اگر داده های واقعی شرایط مفروض مدل را نداشته باشند، استفاده از این روشها ممکن است موجب خطا و گمراهی محقق گردد. حساس بودن بیشتر این مدلها به مشاهدات گمشده و داده های پرت از دیگر محدودیتهای این

[1] : Multivariate regression
[2] : Response variable

روشها به شمار می آید. به علاوه اغلب این روشها قابلیت مدلبندی روابط پیچیده غیرخطی و هچنین اثرات متقابل درجه بالا را ندارند.

داده های با پاسخ چندمتغیره به وفور در پژوهش های متفاوت رخ می دهند اگر توزیع متغیرهای پاسخ مشخص باشد برای تحلیل داده ها از روش های کلاسیک استفاده می شود اما اگر توزیع داده ها مشخص نباشد استفاده از روش های کلاسیک کارایی لازم را ندارد.

در سالهای اخیر برای مدل سازی و پیش بینی پاسخ های آمیخته در آمار کلاسیک روش های متنوعی ارایه شده است که شامل روش های ذیل می باشند:

۱ -برازش مدل های جداگانه برای هر یک از متغیر های پاسخ

۲ -روش های مبتنی بر فاکتورگیری از توزیع توام متغیر های پاسخ [۷-۵]

۳ -روش های مبتنی بر معرفی یک متغیر پنهان برای بیان همبستگی بین متغیر های پاسخ [۱۱-۸]

۴ -روش برآورد معادلات تعمیم یافته [۱۷-۱۵]

با توجه به محدودیت های ذکر شده مدل شبکه عصبی مصنوعی برای تحلیل این گونه داده ها پیشنهاد می شود. در روش شبکه عصبی مصنوعی با توجه به خاصیت یادگیری مدل شبکه عصبی مصنوعی رابطه بین متغیرها از درون خود داده ها کشف می شود و هیچ گونه پیش فرضی در مورد توزیع متغیرهای پاسخ در نظر گرفته نمی شود و همچنین لازم نیست که نوع رابطه ی تابعی، از قبل مشخص باشد و تابع کشف شده می تواند هر نوع تابعی اعم از خطی یا غیر خطی باشد.

شبکه عصبی مصنوعی به طور متنی[1] (محتوایی) داده ها را پردازش می کند، بنابراین شبکه نسبت به داده های پرت و گم شده حساس نیست[۴]. هدف این پژوهش به کارگیری مدل شبکه عصبی مصنوعی برای پیش بینی پاسخ های دو متغیره آمیخته در بیماران قلبی است.

این مطالعه درصدد پیش بینی همزمان کلسترول و سطح LDL کلسترول با استفاده از عوامل خطر مختلف در بیماران قلبی می باشد که با مدل آماری شبکه عصبی مصنوعی انجام خواهد شد.

تا به حال برای بیماری تصلب شرائین[2] عوامل متفاوتی به عنوان عوامل خطر شناخته شده اند که دو مورد از آنها شامل کلسترول خون و سطح LDL[3] کلسترول، به عنوان عامل های خطر ماژور کاملاً شناخته شده، گزارش شده اند. عوامل خطر شناخته شده بیماریهای کلسترول خون و سطح LDL کلسترول، شامل سن، جنس، فشار خون سیستولیک، فشار خون دیاستولیک، نبض، TG، HDL و ... هستند. تاکنون اثر همزمان این عوامل خطر روی کلسترول خون و سطح LDL کلسترول، به صورت آماری کمتر شناخته شده است، بطور مثال ما می دانیم که سن خطر ابتلا به بیماری چربی خون (کلسترول خون) را بالا می برد اما نمی دانیم که آیا سن بالا باعث تغییر در سطح LDL نیز می شود یا این که افزایش چربی خون است که باعث افزایش سطح LDL می شود. لازم به ذکر است که کلسترول خون و سطح LDL کلسترول، دو عامل خطر اصلی شناخته شده بیماری تصلب شرائین بخصوص سکته قلبی هستند، بنابراین عوامل موثر روی کلسترول خون و سطح LDL کلسترول باعث افزایش سکته قلبی

[1] Context
[2] Atherosclerosis disease
[3] Low density chlostrol

خواهند شد. در این پژوهش با استفاده از مدل شبکه عصبی مصنوعی[1]، کلسترول و سطح LDL کلسترول در بیماران قلبی پیش بینی می شود.

۱–۲: اهمیت و ضرورت

در برخی از مطالعات با مجموعه داده هایی مواجه می شویم که دو متغیر پاسخ آمیخته به طور همزمان به عنوان متغیر پاسخ در نظر گرفته می شوند. این حالت زمانی رخ می دهد که برای متغیر پاسخ کمی گروه بندی خاصی به اثبات نرسیده باشد. در این مورد توزیع توام خاصی برای متغیر پاسخ نمی توان در نظر گرفت و همین موضوع تهدید عمده ای برای عدم دقت برآورد پارامترها و پیش بینی با استفاده از روش های کلاسیک است.

مدل شبکه عصبی مصنوعی هیچ گونه پیش فرضی را در مورد داده ها در نظر نمی گیرد و مدل آماری داده ها را با استفاده از اطلاعات درون داده ها برازش می دهد. در واقع نوع رابطه را با استفاده از اطلاعات درون تک تک داده ها کشف می کند[۱].

با توجه به اینکه در پژوهش های گوناگون با داده هایی مواجه می شویم که هیچ گونه اطلاعاتی در مورد توزیع اولیه متغیر پاسخ نداریم برای تحلیل این گونه داده ها به روشی نیاز داریم که توزیع پیشین فرضی را برای داده ها در نظر نگرفته باشد. با توجه به این که در بیشتر موارد از توزیع توام متغیرهای پاسخ اطلاعی وجود ندارد؛ لزوم تحقیق در زمینه هایی مانند شبکه عصبی مصنوعی که در آن نیاز به پیش فرضی بر توزیع متغیرهای پاسخ نیست، امری اجتناب ناپذیر خواهد بود.

[1] neural network model

۱-۲-۱: آمار مرگ و میر ناشی از بیماریهای قلبی عروقی در ایران و جهان

امروزه بیماری های غیرواگیر از جمله بیماری های قلبی عروقی به علت امکانات و تکنولوژی های جدید و تغییر سبک و شیوه زندگی امروزی و روی آوردن به زندگی شهرنشینی و عادات نامناسب غذایی گسترش زیادی پیدا کرده‌اند. طبق گزارش سازمان جهانی بهداشت در سال ۲۰۱۱ مهمترین عوامل خطر ایجاد این بیماری ها در تمام کشورها مربوط به عوامل زیر می‌باشند:

- مصرف دخانیات.
- کلسترول بالا.
- افزایش وزن و چاقی.
- مصرف کم میوه و سبزیجات.
- عدم فعالیت بدنی مناسب.
- افزایش قند خون.
- مصرف داروهای غیرمجاز.

سازمان بهداشت جهانی در سال ۲۰۱۱ اعلام کرد بیماریهای قلبی عروقی، حوادث، سرطان ها مهمترین علل مرگ و میر در جهان می‌باشند. ۲۹/۲٪ از کل مرگ ها در سال ۲۰۱۱ مربوط به بیماری های قلبی عروقی است.

مطالعه بررسی سیمای مرگ و میر در ۲۳ استان کشور در سال ۱۳۸۲ روزانه ۸۰۰ نفر فوت می کنند. بر طبق این گزارش علل عمده مرگ و میر در کشور به ترتیب بیماری های قلبی عروقی، سوانح و حوادث، سرطان ها،

بیماری های تنفسی و بیماری های هنگام تولد می‌باشند. بیماری های قلبی عروقی ۴۶ درصد (۳۶۹ مرگ) این مرگ ها را به خود اختصاص می دهند[۳۵].

در یک مطالعه دیگر با عنوان طرح سلامت و بیماری در ۲۴ استان کشور که در سال ۱۳۷۸ صورت گرفته است سطح کلسترول در ۴۱٪ مردان و ۴۷٪ زنان ۶۰-۴۰ ساله بالا (۲۰۰ mg و بالاتر)گزارش شده است[۳۵].

این آمار نشان دهنده این است که شیوع عوامل خطر بیماری های قلبی عروقی و در نتیجه شیوع مرگ ومیر ناشی از آنها در کشور بالاست و هزینه هنگفتی را به نظام بهداشتی درمانی کشور تحمیل می‌کند، در صورتی که بیماری های قلبی عروقی به سادگی قابل پیشگیری هستند. این مساله لزوم تحقیق در بیماران قلبی را مشخص می کند.

۱-۳: اهداف پژوهش

۱-۳-۱: هدف کلی

پیش بینی همزمان کلسترول خون و سطح LDL کلسترول بیماران قلبی با استفاده از مدل شبکه عصبی مصنوعی

۱-۳-۲: اهداف اختصاصی

۱. پیش‌بینی افراد دارای کلسترول بالاتر از ۲۰۰ در بیماران قلبی، با استفاده از شبکه عصبی مصنوعی

۲. پیش‌بینی سطح LDL در بیماران قلبی با استفاده از شبکه عصبی مصنوعی

۱-۳-۳: اهداف کاربردی

۱-ارایه یک مدل شبکه عصبی مصنوعی برای پیش بینی پاسخ های دو متغیره آمیخته با توجه به عوامل موثر