

**پایان نامه کارشناسی ارشد رشته مهندسی فناوری اطلاعات (M.Sc)**

**گرایش: سیستم های اطلاعاتی**

**عنوان:**

**پیش­بینی طلاق قبل از ازدواج با استفاده از الگوریتم­های داده­کاوی درخت تصمیم**

**بهمن ماه ۹۸**

صفحه تایید هیات داوران

**تقدير و تشكر**

بر خود لازم مي‌دانم تا بعد از حمد و سپاس خداي مهربان، فرصت را غنيمت شمرده و با كلماتي قاصر، از همة تلاشگران عرصة علم و معرفت و اساتيد بزرگواري كه در تمام دوران تحصيل، با تلاش خستگي‌ناپذير و فرمايشات عالمانة خويش، در تعليم و تربيت بنده كوشيده‌اند بخصوص از اساتيد عالي‌مقام، سرکار خانم دکتر سیده صفیه سیادت و جناب آقای دکتر مهدی زاهدی كه با قبول سمت راهنمايي و مشاوره در اين پايان‌نامه، بر من منت گذاشته و با زحمات و راهنمايي‌هاي حكيمانة خويش، مرا در انجام اين مهم ياري فرمودند، كمال تشكر و سپاس را بنمايم. همچنین از مهندس لطفی، مدیر کل نگهداری و مهندسی عملیات سازمان ثبت احوال و مهندس کریمی مدیر دفتر فناوری اطلاعات سازمان ثبت احوال و مهندس اسماعیلی معاونت فناوری ثبت احوال خوزستان که در راستای استخراج داده­ها ی ازدواج و طلاق از پایگاه داده سازمان ثبت احوال یاری رسان اینجانب بوده­اند کمال تشکر را دارم. لازم به ذكر است كه نواقص موجود در اين پژوهش به ناتواني، قلت فهم و سهل‌انگاري بنده مربوط بوده كه پيشاپيش از همة خوانندگان عذرخواهي نموده و ربطي به تلاش‌‌هاي اين بزرگواران ندارد.

**تقديم و اهداء**

**اين دستاورد كوچك را به عنوان هديه‌اي هر چند ناچيز به پدر بزرگوارم به خاطر تمام رنج‌هايش و به مادر عزيزم به خاطر تمام مهرباني‌هايش تقديم نموده و از زحمات و الطاف آنها كمال تقدير و تشكر را دارم.**

فهرست مطالب

[فصل اول کلیات تحقیق 2](#_Toc30374382)

[1-1 مقدمه 3](#_Toc30374383)

[1-2 بیان مسئله 3](#_Toc30374384)

[1-3 اهمیت و ضرورت انجام تحقیق 6](#_Toc30374385)

[1-4 اهداف تحقیق 7](#_Toc30374386)

[1-4-1 هدف اصلی 7](#_Toc30374387)

[1-4-2 اهداف فرعی 7](#_Toc30374388)

[1-5 فرضيه‌ها و سوالات تحقیق 7](#_Toc30374389)

[1-6 ساختار پایان نامه 7](#_Toc30374390)

[فصل دوم: مروری بر مبانی و پیشینه تحقیق 9](#_Toc30374391)

[2-1 مقدمه 10](#_Toc30374393)

[2-2 دادهکاوی 10](#_Toc30374394)

[2-2-1 خدمات قابل ارايه با استفاده دادهكاوي 12](#_Toc30374395)

[2-2-2 فرآیند استخراج دانش در دادهکاوی 13](#_Toc30374396)

[2-2-3 الگوريتمهاي دادهکاوي 14](#_Toc30374397)

[2-3 عناصر دادهکاوی 16](#_Toc30374398)

[2-4 تفاوت دادهکاوی و تحلیل آماری 19](#_Toc30374399)

[2-5 درخت تصمیم 20](#_Toc30374400)

[2-5-1 کاربردها و ویژگیهای درخت تصمیم 21](#_Toc30374401)

[2-5-2 الگوریتم ID3 22](#_Toc30374402)

[2-5-3 الگوریتم C4.5 23](#_Toc30374403)

[2-6 کارهای پیشین 26](#_Toc30374404)

[فصل سوم: معرفی روش پیشنهادی 31](#_Toc30374405)

[3-1 مقدمه 32](#_Toc30374406)

[3-2 مجموعه داده 33](#_Toc30374407)

[3-3 استخراج و پیشپردازش ویژگیهای 33](#_Toc30374408)

[3-3-1 پیشپردازش و آمادهسازی دادهها 34](#_Toc30374409)

[3-3-2 گروهبندی ویژگیهای با تنوع زیاد 36](#_Toc30374410)

[3-4 درخت تصمیم روش پیشنهادی 40](#_Toc30374411)

[3-4-1 نحوه تصمیمگیری در درخت CART 44](#_Toc30374412)

[فصل چهارم: پیاده­سازی و ارزیابی روش پیشنهادی 47](#_Toc30374413)

[4-1 مقدمه 48](#_Toc30374414)

[4-2 دادههای آموزشی 48](#_Toc30374415)

[4-3 ارتباط ویژگیها 49](#_Toc30374416)

[4-4 معیارهای ارزیابی 51](#_Toc30374417)

[4-4-1 منحنی ROC و معیار AUC 53](#_Toc30374418)

[4-4-2 مشخصات سخت افزاری سیستم 54](#_Toc30374419)

[4-5 نتایج ارزیابی 55](#_Toc30374420)

[فصل پنجم: نتیجه­گیری و پیشنهادات آینده 59](#_Toc30374421)

[5-1 مقدمه 60](#_Toc30374422)

[5-2 نتیجه­گیری 60](#_Toc30374423)

5-3 پیشنهادات آینده ....................................................................................................................60

**فهرست جداول**

[جدول (3-1) گروهبندی مشاغل و مدارک بر اساس اطلاعات مرکز آمار ایران 37](#_Toc30374770)

[جدول (3-2) بخشی از مشاغل زیر مجموعه صنعتگر 38](#_Toc30374771)

[جدول (3-3) لیست ویژگیهای استخراج شده برای ایجاد درخت تصمیم 39](#_Toc30374772)

[جدول 4-1: تعریف اختصارات در ارزیابی 52](#_Toc30374773)

[جدول 4-2) مشخصات سخت افزاری سیستم مورد استفاده 54](#_Toc30374774)

[جدول (4-3) نتایج مقایسه درخت تصمیم C4.5 و کارت 57](#_Toc30374775)

فهرست شکل­ها

[شکل ‏2‑1. فرایند کشف دانش در دادهکاوی 14](#_Toc30374977)

[شکل (3-1) مثالی از نحوه هرس کردن درخت 36](#_Toc30374978)

[شکل (3-2) یک نمونه داده ساده با دو ویژگی 41](#_Toc30374979)

[شکل (3-3) زیر درخت تصمیم مربوط به تقسیمبندی شکل (3-2) 42](#_Toc30374980)

[شکل (3-4) تعیین سطح تصمیم برای اختلاف مدرک در حالتی که اختلاف سن کمتر از x1 است 42](#_Toc30374981)

[شکل (3-6) بخشبندی نهایی ویژگی­ها 43](#_Toc30374982)

[شکل (3-7) درخت تصمیم نهایی معادل با شکل (3-6) 44](#_Toc30374983)

[شکل (4-1) نمودار نمایش دو بعد سن زوجین 50](#_Toc30374984)

[شکل (4-2) نمودار ارتباط دو بعد سن زوج و اختلاف سطح سواد زوجین 51](#_Toc30374985)

[شکل (4-3) ماتریس درهم درخت تصمیم C4.5 55](#_Toc30374986)

[شکل (4-4) نمودار ROC مربوط به درخت تصمیم CART 56](#_Toc30374987)

[شکل (4-5) ماتریس درهم درخت تصمیم CART 56](#_Toc30374988)

[شکل (4-6) نمودار ROC درخت تصمیم CART 57](#_Toc30374989)

**چکیده:**

پیش­بینی متغیرها و پدیده­های اجتماعی همواره یکی از دغدغه­های اصلی مدیران برای تصمیم­گیری و سیاست­گذاری در سطح کلان به­شمار می­رود. علم داده­کاوی به بهره­گیری از داده­های واقعی مربوط به پدیده­های مختلف، قابلیت پیش­بنی و استخراج الگوهای نهان از داده­­ها را دارد. در این پایان­نامه یک روش جدید مبتنی بر داده­کاوی برای پیش­بینی طلاق قبل از ازدواج معرفی شده است. برای این منظور از اطلاعات جمع­آوری شده مربوط به دو سال ۹۶ و ۹۷ که شامل آمار ازدواج و طلاق است، استفاده شده است. اطلاعات شامل داده­های خام است که در روش پیشنهادی ابتدا پیش­پردازش­هایی بر روی این داده­ها انجام شده و ویژگی­های جدیدی از ویژگی­های اولیه استخراج می­شوند. سپس گروه­بندی و متوازن­سازی داده­ها انجام می­شود به شکلی که برای دسته­بند مناسب باشند. داده­ها به دو کلاس: تداوم زندگی و طلاق تقسیم­بندی می­شوند که مراحل یادگیری و پیش­بینی توسط الگوریتم درخت تصمیم CART‌ انجام می­شود. نتایج در نهایت بر اساس معیارهای استاندارد، ارزیابی شده و الگوهای مشخصی از درخت تصمیم به منظور پیش­بینی طلاق معرفی شده­اند.

**کلمات کلیدی: پیش­بینی، طلاق، داده­کاوی، درخت تصمیم، الگوریتم CART**

فصل اوّل

# کلیات تحقیق

## مقدمه

پیش­بینی و آینده پژوهی در مسائل اجتماعی، همواره یکی از دغدغه­های اصلی مدیران و کارشناسان برای اخذ تصمیمات راهبردی و صحیح است. پیش­بینی پدیده­ها و متغیرهای اجتماعی، همواره از اهمیت ویژه­ای برای سیاست گذاران و تصمیم­گیران برخوردار است. تصمیم­گیری بر اساس یک پیش­بینی که متکی به حقایق باشد همواره نتیجه­ی بهتری در برابر تصمیم­گیری بر اساس آزمون و خطا دارد. بنابراین پیش­بینی وقایع آینده امری ضروری برای سازمان­ها و نهادهای مختلف اجتماعی، اقتصادی و سیاسی محسوب می­شود به همین دلیل پیش­بینی وقایع آینده مورد توجه محققین در زمینه­های مختلف قرار گرفته است. علارغم توسعه روش­های پیش­بینی، کمتر به کاربرد این روش­ها در پیش­بینی پدیده­هی اجتماعی پرداخته شده است. ازدواج و طلاق به عنوان یکی از مهمترین مسائل اجتماعی به ویژه در جامعه ما به­شمار می­رود که پیش­بینی درصد موفقیت آمیز بودن آن، قبل از ازدواج می­تواند کمک شایانی به کاهش آسیب­های اجتماعی کند. در این پایان نامه به بررسی یک روش مبتنی بر داده­کاوی به منظور پیش­بینی طلاق و میزان امید به زندگی متاهلی بعد از عقد، می­پردازیم [1].

## بیان مسئله

الگوی ازدواج و طلاق به عنوان هسته اجتماع از دیرباز مورد توجه متخصصین، مسئولان و جامعه­شناسان بوده است است. تحولات ایجاد شده در سطح زندگی مردم و افزایش سطح رفاه و زندگی ماشینی، باعث ایجاد تحولات در سطح جامعه و خانواده شده است. تحقیقات گسترده­ای درباره آمار ازدواج و طلاق انجام گرفته است اما کمتر به پیش­بینی این پدیده اجتماعی پرداخته شده است. در این تحقیق از تکنیک­های داده­کاوی برای پیش­بینی طلاق با استفاده از داده­های موجود، استفاده می­کنیم [2].

علم داده­کاوی، ترکیبی از علوم: آمار، کامپیوتر، هوش مصنوعی، یادگیری ماشین، شناسایی الگو و بازنمایی بصری داده است. داده­کاوی فرایندی به منظور استخراج الگوها و مدل­های نهان از داده­های خام در سطح نسبتاْ وسیع است به شکلی که این الگوها توسط انسان قابل درک باشند. در حالت عادی این الگوها توسط روش­های قدیمی و برنامه­نویسی ساده، قابل شناسایی نیستند [3]. در داده­کاوی با بهره­گیری از داده­های واقعی و شناسایی تعدادی ویژگی برای داده­ها، سعی می­شود توسط تکنیک­های یادگیری، ارتباطات نهان بین ویژگی­ها و کلاس داده­ها استخراج شود. یکی از زمینه­های مورد بحث در داده­کاوی، پیش­بینی است. پیش­بینی و تشخیص اینکه یک ازدواج چقدر امکان دارد به طلاق منجر شود، تابحال توسط تکنیک­های داده­کاوی به صورت گسترده انجام نگرفته است. اما قابلیت­های داده­کاوی و به­یژه درخت تصمیم این امکان را برای ما فراهم می­کند که با شناسایی تعدادی ویژه مهم زوجین امکان وقوع طلاق را بررسی کرد [4].

ما در این تحقیق از داده­های خام ازدواج که در سال­های 96 و 97 در ایران ثبت شده و داده­های طلاق مربوط به ازدواج­های این دوسال، به پیش­بینی طلاق با استفاده از تکنیک­های داده­کاوی برای داده­های جدید می­پردازیم. ویژگی­های این داده­ها تاریخ تولد زوجین (که خود شامل ویژگی­های سن زوجین هنگام ازدواج، اختلاف سنی زوجین، ماه تولد زوجین می­باشد)، شغل زوجین، تحصیلات زوجین، دفعات ازدواج و طلاق زوجین، تاریخ ثبت ازدواج و در صورت منجر شدن به طلاق تاریخ طلاق می­باشد.

فرآیند حل یک مسئله توسط داده­کاوی، شامل مراحل پاک­سازی و پیش­پردازش داده، یکپارچه­سازی داده­ها، انتخاب ویژگی­ها، تبدیل داده، اعمال الگوریتم داده­کاوی و ارزیابی الگو است. داده­کاوی شامل تکنیک­هایی جهت پیش­بینی بر پایه­ی آمار، رگرسیون، قوانین انجمنی، پردازش زبان طبیعی، شبکه عصبی و سایر روش­های یادگیری ماشین است هر کدام از این الگوریتم­ها با توجه به هدف نهایی داده­کاوی و همچنین نوع مدل­سازی داده­ها قابل استفاده هستند.

درخت تصمیم یکی از الگوریتم­های مشهور در زمینه داده­کاوی است. درخت تصمیم یک روش یادگیری باناظر محسوب می­شود که در آن باید داده­های آموزشی شامل برچسب کلاس مودر نظر باشند. هدف اصلی در درخت تصمیم­گیری، تقسیم داده­ها به صورت بازگشتی به زیر مجموعه­هایی است که در آن هر زیرمجموعه شامل وضعیت همگنی از متغیر هدف است. در هر زیردرخت، تمام ویژگی­های ورودی به منظور تآثیرشان بر روی ویژگی­های قابل پیش­بینی به شکل بازگشتی ارزیابی می­شوند. [5].

روش یادگیری درخت تصمیم یک روش یادگیری ساده، سریع و با قدرت تفسیرپذیری بالا است. در عین حال دارای قدرت تفکی­پذیری بالا است. به ویژه برای داده­هایی با ویژگی­های گسسته­ی کیفی این روش دارای عملکرد بسیار خوبی است. روش­های مختلفی برای ایجاد درخت تصمیم وجود دارد و می توان از الگوریتم­های متنوعی برای تعیین چگونگی ­بخش­بندی درخت استفاده کرد. هر مسیر در درخت از ریشه تا برگ، بیان­کننده­ی یک قانون و الگوی استخراج شده است. درخت تصمیم یک الگوریتم ایجاد مدل طبقه­بندی[[1]](#footnote-2) و رگرسیون است. این روش طبقه­بندی در مدل­های پیش­بینی درمورد ویژگی­های پیوشته و گسسته فراهم شده است. برای ویژگی­های گسسته، درخت تصمیم پیش­بینی­هایی را بر اساس روابط بین ستون­های ورودی در یک مجموعه انجام می­دهد. از مقادیر ویژگی­های مربوط به این ستون­ها برای پیش­بینی مقدار (در رگرسیون) یا کلاس (طبقه­بندی) مربوط به ستون هدف[[2]](#footnote-3) استفاده می­شود. ستون هدف، ستونی است که برچسب مربوط به داده را شامل می­شود که توسط فرد خبره برای آن در نظر گرفته شده است. برای ویژگی­های پیوسته الگوریتم از روش رگرسیون خطی برای مشخص­کردن اینکه درخت تصمیم در کجا بخش­بندی شود، استفاده می­نماید. اگر بیش از یک ویژگی هدف قابل پیش­بینی موجود باشد یا اگر داده­های اولیه شامل یک جدول تودرتو باشند، در این حالت، یک مجموعه­ی قابل پیش­بینی وجود دارد و الگوریتم برای هر یک از ویژگی­های قابل پیش­بینی یک درخت تصمیم جداگانه ایجاد می­کند. در این تحقیق قصد داریم ار درخت تصمیم به عنوان یکی از روش­های داده کاوی و یادگیری با­نظارت جهت پیش­بینی استفاده نماییم.

## اهمیت و ضرورت انجام تحقیق

طلاق به عنوان یکی از مهمترین آسیب­های اجتماعی، تاثیر بسیار زیادی بر پیکره­ی اجتماع وارد می­کند و آثار بسیار مخربی بر جامعه ایجاد می­کند. طلاق دارای علل بسیار گسترده­ای است که از دخالت­های نابجای اطرافیان گرفته تا مشکلات معیشتی و سو رفتارهای زوجین را در بر می­گیرد. حتی گاهی علت اصلی طلاق در پس عدم تفاهمات و مشکلات پیش پا افتاده مخفی می­شود تنها به این علت که زوجین از بیان آن اکراه دارند و یا اینکه حتی در بعضی موارد خودشان هم دلیل اصلی طلاق را نمی­دانند. وجود یک روش (نه لزوماْ با دقت صد درصد) که بتواند احتمال تداوم زندگی زوجین را پیش­بینی کند، می­تواند از اهمیت بالایی برخوردار باشد. علم داده­کاوی با بهتره­گیری از داده­های واقعی می­تواند الگوهای نهفته از آن­ها استخراج کند و وقوع یک پدیده را در آیند با دقت مناسبی پیش­بینی کند. بنابراین وجود یک سیستم که بتواند داده­های گذشته مربوط به ازدواج و طلاق را بررسی کرده و احتمال طلاق یا تداوم زندگی را پیش­بینی کند بسیار لازم و ضروری است [۶].

## اهداف تحقیق

### هدف اصلی

پیش­بینی طلاق قبل از ازدواج با استفاده از درخت تصمیم که یکی از الگوریتم­های داده­کاوی می­باشد و کمک به شناسایی ازدواج­های صحیح­تر

### اهداف فرعی

* نتایج پیش بینی در اختیار سیاست گذاران و تصمیم گیران قرار گیرد.
* هر کدام از متغیرها چقدر بر میزان ثبات ازدواج تأثیر دارد.

## فرضيه‌­ها و سوالات تحقیق

* راهکارهای درخت تصمیم در پیش­بینی طلاق کدامند؟
* راه­حل­های تشخیص ازدواج موفق یا ناموفق با استفاده از درختان تصمیم کدامند؟
* میزان تأثیر هر کدام از متغیر­ها در موفق یا ناموفق بودن ازدواج چقدر است؟

## ساختار پایان نامه

این پایان نامه شامل پنج فصل است. فصل اول شامل کلیات تحقیق یعنی؛ بیان مسئله، ضرورت و اهمیت تحقیق و اهداف و فرضیات است. در فصل دوم مبانی و پیشینه­ی تحقیق بیان می­شود و به بررسی کارهای انجام شده در این زمینه و معایب و مزایای آن می­پردازد. بعد از بررسی کارهای انجام شده، فصل سوم به معرفی جزئیات روش پیشنهادی می­پردازد. فصل چهارم شامل شبیه­سازی روش پیشنهادی در یک محیط نرم­افزاری و ارزیابی و مقایسه نتایج بدست آمده با کارهای قبلی است. در فصل پنجم نتایج نهایی تحقیق بیان می­شود و راهکارهایی برای ادامه تحقیق و بهبود آن در آینده معرفی می­شود.

فصل دوم

# مروری بر مبانی و پیشینه تحقیق

* فصل دوم: مروری بر مبانی تحقیق و کارهای پیشین

## مقدمه

با پیدایش عصر کامپیوتر و فناوری اطلاعات، داده­ها نقش بسیار پررنگی در زندگی روزمره ما پیدا کرده­اند. بسیاری از سازمان­ها، از این داده­ها به عنوان یک سرمایه بزرگ یاد می­کنند که می­توانند در زمینه­های مختلف از آن­ها استفاده کنند. زمانی که با استفاده از قواعد موجود، داده­های دیگری از این داده­های خام[[3]](#footnote-4) اولیه استخراج می­کنیم، به داده­های جدید، اطلاعات[[4]](#footnote-5) می­گوییم. اما داده­کاوی در یک سطح بالاتر داده­ها را تحلیل می­کند و الگوهای ناشناخته را از آن­ها استخراج می­کند که به آن دانش[[5]](#footnote-6) می­گوییم. تحلیل این دانش استخراج شده می­تواند در بسیاری از زمینه­های مرتبط با داده­ها مفید باشد. در این تحقیق سعی می­کنیم وقوع طلاق را برای یک ازدواج بر اساس تکنیک­های داده­کاوی پیش­بینی کنیم. به همین دلیل در این فصل به معرفی مبانی علم داده­کاوی و تکنیک­های آن می­پردازیم و در ادامه تعدادی از کارهای انجام شده بررسی می­شوند [7].

## داده­کاوی

داده­کاوی[[6]](#footnote-7) عبارت است از فرایند یافتن اطلاعات نهفته و استخراج الگوها و روابط مشخص بین ویژگی­های مختلف داده­ها در حجم انبوه. نیاز اصلی داده­کاوی وجود داده­های خام واقعی است. هرچه دقت داده­ها بیشتر باشد و ویژگی­های استخراج شده ارزش بیشتری داشته­ باشند، تکنیک داده­کاوی به همان میزان دقیق­تر خواهد بود. بنابراین لازمه­ی یک الگوریتم داده­کاوی در هر مسئله­ای وجود داده­های واقعی برای آن است. در این روش، شما با بهره‌گیری از الگوریتم­های‌های مختلف می‌توانید اطلاعات استخراج شده را برای دست­یافتن به اهدافی مانند: کاهش هزینه‌ها، افزایش درآمد، بهبود ارتباط با مشتری، کاهش ریسک و غیره مورد استفاده قرار دهید [13].

1. Classification [↑](#footnote-ref-2)
2. Target [↑](#footnote-ref-3)
3. Data [↑](#footnote-ref-4)
4. Information [↑](#footnote-ref-5)
5. Knowldge [↑](#footnote-ref-6)
6. Data Mining [↑](#footnote-ref-7)