چکیده

بتن به عنوان پر م صرف ترین ماده در ساخت سازه‌های مختلف، از اهمیت فنی و اقت صادی زیادی برخوردار می‌باشد. برای تولید بتن روش‌های استاندارد گوناگونی معرفی شده که نیازمند انجام آزمایش‌های پر هزینه و زمان‌بر هستند. به منظور کاهش زمان و هزینه‌های فوق الذکر، در کنار روش‌های مرسوم و سنتی، می‌توان از قابلیت‌های علم داده مانند الگوریتم‌های یادگیری ما شین برای پیش بینی خواص گوناگون م صالح بر ا ساس داده های موجود بهره گرفت. در این پژوهش روابط بین متغیر های مؤثر در تولید بتن و پیش بینی خواص بتن با اســتفاده از علم داده، مورد بررســی قرار گرفته اســت. بدین منظور با جمع‌آوری داده‌های مختلف انواع طرح‌های مخلوط بتن و ا صلاح مجموعه دادها، از الگوریتم‌های یادگیری ما شین برای پیش بینی خواص بتن ا ستفاده گردید. نتایج کلی حا صل که صرفاً بر پایه آ شکار سازی اطلاعات در الگوهای پنهان داده‌ها بد ست آمدند، توافقی کاملی با حقایق موجود در فناوری بتن دارند. با افزایش مقدار نســبت آب به مواد ســیمانی، مقاومت بتن کاهش می‌یابد. نســبت مقاومت کشــشــی به مقاومت فشــاری بتن تقریباً برابر ۰/۰۸ می‌باشــد. هم‌چنین با افزایش مقاومت فشــاری بتن این نســبت کاهش پیدا می‌کند. به عنوان نمونه ای دیگر این‌که با افزایش سن و عیار سیمان در مخلوط بتن، مقاومت آن افزایش می‌یابد.

هدف این تحقیق نمایش توانمندی علم داده برای کاربرد در مطالعه و بررســی مصــالح بتنی اســت. در این پژوهش از روش‌های مختلف یادگیری ماشــین برای پیش بینی خواص بتن اســتفاده شــده اســت و نتایج به دست آمده برای اهداف طراحی سازه‌های بتنی قابل قبول هستند. با گسترش بانک داده‌ها می‌توان این روش را توسعه داد و در شاخه‌های مختلف مهندسی عمران از آن استفاده نمود.

واژه‌های کلیدی: علم داده، الگوریتم های یادگیری ماشــین، خواص مکانیکی بتن، مجموعه داده‌ی انواع مخلوط بتن

فهرست مطالب

فصل ۱: مقدمه **۱**
۱-۱- عنوان تحقیق ۲
۱-۲- بیان مسأله ۲
۱-۳- ضرورت تحقیق ۲
۱-۴- هدف ۳
۱-۵- فصول پژوهش ۳

فصل ۲: مروری بر ادبیات موضوع **۵**
۲-۱- مقدمه ۶
۲-۲- تاریخچه موضوع ۶
۲-۳- تعریف علم داده ۸
۲-۴- متخصص علم داده ۹
۲-۵- تاریخچه اهمیت یافتن داده ۱۰
۲-۶- یادگیری ماشین ۱۲
۲-۷- انواع یادگیری ماشین ۱۵
۲-۷-۱- یادگیری نظارت‌شده ۱۵
۲-۷-۲- یادگیری نظارت‌نشده ۱۶
۲-۷-۳- یادگیری تقویتی ۱۷
۲-۷-۴- یادگیری عمیق ۱۷
۲-۸- روند تکامل انواع روش‌های یادگیری ماشین ۱۸
۲-۹- اهمیت آمار در یادگیری ماشین ۲۰
۲-۱۰- مراحل پیاده سازی یادگیری ماشین ۲۰
۲-۱۰-۱- تمیزکاری داده ۲۰
۲-۱۰-۲- تشخیص نوع داده ۲۱
۲-۱۰-۳- مدل‌سازی مبتنی بر یادگیری ماشین ۲۱
۲-۱۰-۴- ارزیابی مدل ساخته شده ۲۲
۲-۱۰-۵- کاربرد مدل‌های مبتنی بر یادگیری ماشین برای پیش بینی ۲۳
۲-۱۰-۶- استفاده از یادگیری ماشین در علم عمران ۲۳

فصل ۳: روش تحقیق **۲۵**
۳-۱- مقدمه ۲۶
۳-۲- نحوه جمع‌آوری داده ۲۷
۳-۳- مشکلات جمع‌آوری داده ۲۸
۳-۴- مجموعه داده اصلی ۲۹

۳-۵- نرم‌افزارهای یادگیری ماشین ۳۱
۳-۶- نرم‌افزار منتخب برای اجرای مدل‌سازی ۳۳
۳-۷- الگوریتم‌های مدل‌سازی ۳۷
۳-۱- انواع الگوهای مدل‌سازی ۳۸

فصل ۴: نتایج و تفسیر آن‌ها **۴۲**
۴-۱- مقدمه ۴۳
۴-۲- الگوی شماره‌ی ۱ ۴۴
۴-۲-۱- نمودار پراکندگی ۴۴
۴-۲-۲- برازش خطی ۴۸
۴-۲-۳- همبستگی متغیرها ۴۹
۴-۲-۴- شبکه عصبی ۴۹
۴-۲-۵- دسته‌بندی مقاومت فشاری ۵۱
۴-۲-۶- درخت تصمیم ۵۲
۴-۳- الگوی شماره ۲ ۵۴
۴-۳-۲- نمودار پراکندگی ۵۴
4-3-3- برازش خطی ۵۷
۴-۳-۴- همبستگی متغیرها ۵۸
۴-۳-۵- شبکه عصبی ۵۸
۴-۳-۶- درخت تصمیم ۶۰
۴-۴- الگوی شماره ۳ ۶۱
۴-۴-۱- نمودار پراکندگی ۶۲
۴-۴-۲- برازش خطی ۶۵
۴-۴-۳- همبستگی متغیرها ۶۶
۴-۴-۴- شبکه عصبی ۶۶
۴-۴-۵- درخت تصمیم ۶۸
۴-۵- الگوی شماره ۴ ۷۰
۴-۵-۱- نمودار پراکندگی ۷۱
۴-۵-۲- برازش خطی ۷۲
۴-۵-۳- همبستگی متغیرها ۷۲
۴-۵-۴- شبکه عصبی ۷۳
۴-۵-۵- درخت تصمیم ۷۶
۴-۶- الگوی شماره ۵ ۷۷
۴-۶-۱- نمودار پراکندگی ۷۸
۴-۶-۲- خوشه‌بندی ۷۹
۴-۶-۳- برازش خطی ۸۱

۴-۶-۴- همبستگی متغیرها ۸۱
۴-۶-۵- شبکه عصبی ۸۲
۴-۶-۶- درخت تصمیم ۸۳
۴-۶-۷- الگوی شماره ۶ ۸۶
۴-۶-۸- نمودار پراکندگی ۸۶
۴-۶-۹- خوشه‌بندی ۸۹
۴-۶-۱۰- برازش خطی ۹۰
۴-۶-۱۱- همبستگی متغیرها ۹۱
۴-۶-۱۲- شبکه عصبی ۹۱
۴-۶-۱۳- درخت تصمیم ۹۳
۴-۷- الگوی شماره ۷ ۹۵
۴-۷-۱- شبکه عصبی برای پیش‌بینی مقدار نسبت آب به سیمان ۹۶
۴-۷-۲- شبکه عصبی برای پیش‌بینی مقدار سیمان ۹۷
۴-۷-۳- شبکه عصبی برای پیش‌بینی مقدار ماسه ۹۹
۴-۷-۴- شبکه عصبی برای پیش‌بینی مقدار شن ۱۰۰
۴-۷-۵- پیش‌بینی نهایی ۱۰۲
۴-۸- الگوی شماره ۸ ۱۰۳
۴-۸-۱- رابطه مقاومت فشاری و مقاومت کششی ۱۰۳
۴-۸-۲- رابطه مقاومت فشاری و ضریب ارتجاعی ۱۰۴

فصل ۵: جمع‌بندی و پیشنهادها **۱۰۶**

۵-۱- مقدمه ۱۰۷
۵-۲- جمع‌بندی ۱۰۷
۵-۳- پیشنهادها ۱۰۹

مراجع **۱۱۰**

فهرست اشکال

شکل (۳-۱) جریان کاری اجرای پژوهش..... ۲۶
شکل) ۳-۲(دو نما از قسمتی از بانک داده اصلی..... ۳۰
شکل (۳-۳) رابط کاربری نرم‌افزار مورداستفاده..... ۳۴
شکل) ۳-۴(مجموعه عملگرهای اصلی نرم افزار RapidMiner..... ۳۵
شکل (۳-۵) نمونه ای از متغیرهای در دسترس برای الگوریتم شبکه عصبی..... ۳۶
شکل) ۳-۶(محیط اسکریپت نویسی در برنامه RapidMiner..... ۳۷
شکل (۴-۱) متغیرهای الگوی شماره ۱..... ۴۴
شکل (۴-۲) نمودار نسبت آب به سیمان و مقاومت فشاری بتن..... ۴۵
شکل (۴-۳) نمودار نسبت آب به سیمان و مقاومت فشاری بتن برای بازه بیشتر از ۰/۳۵..... ۴۶
شکل (۴-۴) نمودار نسبت آب به سیمان و مقاومت فشاری بتن برای بازه کمتر از ۰/۳۵..... ۴۷
شکل (۴-۵) معماری شبکه عصبی الگوی شماره ۱..... ۵۰
شکل (۴-۶) درخت تصمیم الگوی شماره ۱..... ۵۲
شکل (۴-۷) نمودار مربوطه در مقاله برای ارزیابی اثر سن و نسبت آب به سیمان در مقاومت فشاری[51].۵۳
شکل (۴-۸) متغیرهای الگوی شماره ۲..... ۵۴
شکل (۴-۹) نمودار نسبت آب به سیمان به ضریب ارتجاعی..... ۵۴
شکل (۴-۱۰) نمودار چگالی بتن تازه و ضریب ارتجاعی..... ۵۵
شکل) ۴-۱۱(نمودار نسبت چگالی بتن تازه به ضریب ارتجاعی برای مقادیر چگالی بیشتر ۲۱۰۰ کیلوگرم بر مترمکعب..... ۵۶
شکل (۴-۱۲) معماری شبکه عصبی مناسب برای الگوی شماره‌ی ۲..... ۵۹
شکل (۴-۱۳) درخت تصمیم الگوی شماره‌ی ۲..... ۶۰
شکل (۴-۱۴) متغیرهای الگوی شماره ۳..... ۶۱
شکل (۴-۱۵) نمودار نسبت آب به سیمان و مقاومت کششی در الگوی شماره‌ی ۳..... ۶۲
شکل (۴-۱۶) نمودار چگالی ویژه بتن تازه و مقاومت کششی..... ۶۳
شکل) ۴-۱۷(نمودار چگالی بتن تازه و مقدار مقاومت کششی برای محدوده چگالی بیش از ۲۱۰۰ کیلوگرم بر مترمکعب..... ۶۴
شکل (۴-۱۸) معماری شبکه عصبی الگوی شماره۳..... ۶۷
شکل (۴-۱۹) درخت تصمیم الگوی شماره ۳..... ۶۹
شکل (۴-۲۰) متغیرهای الگوی شماره ۴..... ۷۰
شکل (۴-۲۱) نمودار نسبت آب به سیمان و اسلامپ..... ۷۱
شکل (۴-۲۲) معماری مناسب شبکه عصبی الگوی شماره ۴..... ۷۴

شکل (۴-۲۳) معماری شبکه عصبی الگوی شماره ۴ با روان کننده..... ۷۵
شکل (۴-۲۴) متغیرهای الگوی شماره ۵..... ۷۷
شکل (۴-۲۵) نمودار ماده فوق روان کننده و مقاومت فشاری..... ۷۸
شکل (۴-۲۶) ساختار خوشه‌بندی الگوی شماره ۵..... ۷۹
شکل (۴-۲۷) خوشه‌بندی داده الگوی شماره۵..... ۸۰
شکل (۴-۲۸) معماری مناسب شبکه عصبی برای الگوی شماره ۵..... ۸۲
شکل (۴-۲۹) درخت تصمیم الگوی شماره ۵..... ۸۴
شکل (۴-۳۰) متغیرهای الگوی شماره ۶..... ۸۶
شکل (۴-۳۱) نمودار نسبت ماسه به شن و مقاومت فشاری..... ۸۷
شکل (۴-۳۲) نمودار دسته اول نسبت ماسه به شن و مقاومت فشاری..... ۸۸
شکل (۴-۳۳) ساختار خوشه‌بندی الگوی شماره ۶..... ۸۹
شکل (۴-۳۴) خوشه‌بندی الگوی شماره ۶..... ۸۹
شکل (۴-۳۵) معماری مناسب شبکه عصبی الگوی شماره ۶..... ۹۲
شکل (۴-۳۶) درخت تصمیم الگوی شماره ۶..... ۹۴
شکل (۴-۳۷) متغیرهای الگوی شماره ۷..... ۹۵
شکل (۴-۳۸) معماری مناسب شبکه عصبی الگوی شماره ۷ پیش‌بینی نسبت آب به سیمان..... ۹۷
شکل (۴-۳۹) معماری مناسب شبکه عصبی الگوی شماره ۷ پیش‌بینی سیمان..... ۹۸
شکل (۴-۴۰) معماری مناسب شبکه عصبی الگوی شماره ۷ پیش‌بینی مقدار ماسه..... ۱۰۰
شکل (۴-۴۱) معماری مناسب شبکه عصبی الگوی شماره ۷ پیش‌بینی مقدار شن..... ۱۰۱
شکل (۴-۴۲) نمودار مقاومت کششی برحسب مقاومت فشاری..... ۱۰۳

فهرست جداول

جدول (۴-۱) جدول علائم اختصاری..... ۴۳
جدول (۴-۲) همبستگی متغیرهای الگوی شماره ۱..... ۴۹
جدول (۴-۳) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۱..... ۴۹
جدول (۴-۴) جدول ورودی‌های الگوی شماره ۱..... ۵۰
جدول (۴-۵) نتیجه پیش‌بینی شبکه عصبی..... ۵۱
جدول (۴-۶) جدول همبستگی متغیرهای الگوی شماره ۲..... ۵۸
جدول (۴-۷) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۲..... ۵۸
جدول (۴-۸) جدول داده موردنظر برای پیش‌بینی مقدار ضریب ارتجاعی..... ۵۹
جدول) ۴-۹(مقدار پیش‌بینی‌شده ضریب ارتجاعی..... ۵۹
جدول (۴-۱۰) همبستگی متغیرهای الگوی شماره‌ی ۳..... ۶۶
جدول (۴-۱۱) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۳..... ۶۷
جدول (۴-۱۲) جدول ورودی پیش‌بینی شبکه عصبی الگوی شماره ۳..... ۶۸
جدول (۴-۱۳) پیش‌بینی شبکه عصبی الگوی شمار ۳..... ۶۸
جدول (۴-۱۴) همبستگی متغیرهای الگوی شماره ۴..... ۷۳
جدول (۴-۱۵) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۳..... ۷۳
جدول (۴-۱۶) ورودی شبکه عصبی الگوی شماره ۴..... ۷۴
جدول (۴-۱۷) پیش‌بینی مقدار اسلامپ شبکه عصبی..... ۷۴
جدول (۴-۱۸) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۴ با روان کننده..... ۷۵
جدول (۴-۱۹) ورودی شبکه عصبی الگوی شماره ۴ با روان کننده..... ۷۶
جدول (۴-۲۰) جدول پیش‌بینی شبکه عصبی..... ۷۶
جدول (۴-۲۱) همبستگی متغیرهای الگوی شماره ۵..... ۸۱
جدول (۴-۲۲) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۵..... ۸۲
جدول (۴-۲۳) جدول ورودی شبکه عصبی الگوی شماره ۵..... ۸۳
جدول (۴-۲۴) جدول پیش‌بینی شبکه عصبی الگوی شماره ۵..... ۸۳
جدول (۴-۲۵) جدول همبستگی متغیرهای الگوی شماره ۶..... ۹۱
جدول (۴-۲۶) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۶..... ۹۲
جدول (۴-۲۷) جدول ورودی پیش‌بینی الگوی شماره ۶..... ۹۳
جدول) ۴-۲۸(جدول نتیجه پیش‌بینی شبکه عصبی الگوی ۶..... ۹۳
جدول (۴-۲۹) جدول اولیه الگوی معکوس شماره ۷..... ۹۶
جدول (۴-۳۰) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۷ پیش‌بینی نسبت آب به سیمان ۹۶

جدول (۴-۳۱) جدول پیش‌بینی مرحله اول نسبت به آب به سیمان..... ۹۷
جدول (۴-۳۲) جدول اولیه پیش‌بینی مقدار سیمان در الگوی شماره ۷..... ۹۷
جدول (۴-۳۳) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۷ پیش‌بینی مقدار سیمان..... ۹۸
جدول (۴-۳۴) جدول پیش‌بینی مقدار سیمان الگوی شماره ۷..... ۹۸
جدول (۴-۳۵) جدول اولیه پیش‌بینی مقدار ماسه در الگوی شماره ۷..... ۹۹
جدول (۴-۳۶) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۷ پیش‌بینی مقدار ماسه..... ۹۹
جدول (۴-۳۷) مقدار پیش‌بینی‌شده برای ماسه در الگوی شماره ۷..... ۱۰۰
جدول (۴-۳۸) جدول اولیه پیش‌بینی مقدار شن در الگوی شماره ۷..... ۱۰۱
جدول (۴-۳۹) جدول تحلیل حساسیت شبکه عصبی الگوی شماره ۷ پیش‌بینی مقدار شن..... ۱۰۱
جدول (۴-۴۰) مقدار پیش‌بینی‌شده برای شن در الگوی شماره ۷..... ۱۰۲
جدول (۴-۴۱) جدول نهایی الگوی شماره ۷..... ۱۰۲

فصل ۱:

مقدمه

۱-۱- عنوان تحقیق

در این پژوهش ســـعی بر این اســـت که برخی خواص مخلوط‌های بتنی را که می‌توان با انجام آزمایش در آزمایـ شگاه مورد برر سی قرار داد با ا ستفاده الگوریتم های یادگیری ما شین ا ستخراج کرد. این امر برر سی مخلوط‌های بتنی را بسیار تسهیل می‌بخشد.

۱-۲- بیان مسأله

آزمایش‌های بتنی در آزمایشگاه بدین صورت بوده است که در آن پژوهشگر با ساختن نمونه از طرح مخلوط مورد نظر خود و اجرای آزمایش مربوطه ویژگی های طرح مخلوط رو مورد بررسی قرار می‌دهد. اما پژوهشگر می‌تواند با اســـتفاده از داده‌ی موجود از آزمایش‌های قبلی خود یا دیگر پژوهشـــگران اســـتفاده کند و با پیاده سازی الگوریتم منا سب از مجموعه الگوریتم های موجود در هوش م صنوعی، یادگیری ما شین را روی داده اجرا کنده و مدل یادگیری شـــده ای بســـازد. ســـپس از مدل مبتنی بر یادگیری خود نتایج مورد نیاز خود را به‌دست آورد.

۱-۳- ضرورت تحقیق

در پروژه‌های عمرانی بنا به محا سبات صورت گرفته نیاز به بتن با م شخ صات فیزیکی خا صی ا ست که طبق این م شخ صات طرح مخلوط بتن به د ست می‌آید. سپس طرح مخلوط به د ست آمده در آزمایـ شگاه مورد آزمایش قرارگرفته و صحت محا سبات سنجیده می شود و در نهایت در پروژه مورد نظر مورد ا ستفاده قرار می‌گیرد. اما این فرآیند، زمان بر و هزینه بر ا ست. در صورتی که نیاز با شد نتایج آزمایش در زمان اندکی به دســـت بیاید یا شـــرایط آزمایش نباشـــد راه حل چیســـت؟ در برخی موارد مهندســـین به دنبال بتن با ویژگی خا صی هستند که تا کنون ساخته نشده ا ست، یافتن طرح مخلوط بتن با ویژگی مورد نظر نیازمند آزمون و خطا و آزمایش مکرر نمونه‌های مختلف مخلوط بتنی اســـت. حال در صـــورتی که شـــرایط برای آزمون و خطا فراهم نباشد و یا محدودیت‌های برای آزمایش وجود داشته باشد این امر نیست با بن بست روبرو می‌شود، از این رو در این پژوهش راه حل مبتنی بر یادگیری ماشین در سیستم های کامپیوتری ارائه می‌شود که موارد

فوق و بسیاری از مشکلات مشابه راه تا حد زیادی حل خواهد کرد.

۱-۴- هدف

هدف اصلی این پژوهش این است که آزمایش‌هایی که به صورت متداول در آزمایشگاه‌ها اجرا می‌شده است را به صورت یادگیری ماشین مبتنی بر هوش مصنوعی اجرا کند و در نهایت نتایج حاصل را با روش‌های سنتی آزمایشگاهی مقایسه کند. پژوهشگر در این پژوهش قصد دارد که تا دستاورد های واضح و متداولی که در علم طرح مخلوط بتن تا کنون وجود داشته است را با یادگیری ماشین اثبات کند و نشان دهد که دستاوردهای حاصل از تحلیل های مبتنی بر هوش مصنوعی همان مفاهیم را عرضه می‌کند و این خود نمایش صحت این روش نوین در طرح مخلوط بتن است. پژوهشگر در این فرآیند فقط دسترسی به داده دارد و با استفاده از داده اطلاعات را استخراج خواهد کرد. از سوی دیگر در این پژوهش بر اساس الگوریتم‌های مناسب مدلسازی مناسب برای به دست آوردن تحلیل یا پیش بینی صحیح از خوص طرح مخلوط ارائه می‌گردد تا دیگر پژوهشگران با تکیه بر یادگیری ماشین و الگوریتم‌های هوش مصنوعی درکنار نوآوری های خود دستاوردهای جدیدی در علم طرح مخلوط بتن به رقم بزنند. به کار گیری علم یادگیری ماشین موجب می‌شود که در نهایت باعث به وجود امدن تحول بزرگی در مهندسی عمران و علم طرح مخلوط بتن خواهد شد. دیگر مهندسان نیاز به طرح مخلوط یا طراحی بتن با ویژگی مورد نظر ندارد و می‌توانند با فقط دانستن ویژگی های مورد نظر بتن طرح مخلوط خود را از سیستم یادگیری شده به دست بیاورند.

۱-۵- فصول پژوهش

پژوهش پیش رو متمرکز بر استفاده از هوش مصنوعی و الگوریتم‌های یادگیری ماشین برای پیش‌بینی برخی از خواص مخلوط‌های بتنی است. در فصل دوم خلاصه‌ای از پیشینه موضوع و تاریخچه‌ای از علم داده و یادگیری ماشین بیان خواهد شد. در این فصل نحوه و دلیل ایجاد علم داده ذکر می‌گردد و متخصص علم داده توصیف می‌شود. هم‌چنین مفاهیم علم داده و روش‌های پیاده سازی آن بیان می‌شود.در ادامه فصل دوم روش‌های یادگیری ماشین نیز به‌صورت موجز نیز اشاره می‌شود.

در فصل سوم روش کار بیان خواهد شد. در این فصل روش جمع آوری داده و مشکلاتی پژوهشگر که در این

فرآیند روبرو شد ذگر می‌گردد. نرم افزارهای موجود در زمینه پژوهش مقایسه می‌گردد و نرم افزار منتخب و مناسب معرفی می‌شود. سپس اصول مدلسازی که ذکر می‌شود. برای مدلسازی متغیرها گروه بندی مشخص می شوند که نتیجه این امر ایجاد ۸ سناریو مدل‌سازی خواهد بود. در این سناریو با ا ستفاده از الگوریتم‌های مشخص مدلسازی اجرا می‌شود.

در فصل چهارم نتایج به دست آمده از مدل‌سازی‌های اجرا شده ذکر خواهد گردید. در این فصل تمامی نتایج اعم از نمودار، جدول، شکل و رابطه جبری درج و در هر قسمت توضیح کمی و کیفی مربوطه نیز بیان خواهد شد. هم‌چنین نتایج به دست آمده با مقالات یا پژوهش‌های مرتبط مقایسه می‌شوند.

و درنهایت در فصل پنجم نتیجه‌گیری کلی و ادامه راه برای پژوهشگران به‌صورت خلاصه ذکر خواهد شد.

فصل ۲:

مروری بر ادبیات موضوع

۲-۱- مقدمه

هدف از این فصل مروری بر پیشینه تحقیق ، بررسی و طبقه‌بندی یافته‌های تحقیقات دیگر محققان در سطح دنیا و تعیین و شناسایی خلاء های تحقیقاتی است.

۲-۲- تاریخچه موضوع

هر چند در بدو امر به نظر می رسد علم‌داده دانش جدیدی است، اما شواهد تاریخی نشان می دهد این علم پیشینه ای دیرینه داشته و وجود ضمنی آن به سال‌های بسیار دور و حتی به دوران قبل از ناپلئون بناپارت باز می گردد، چنان که شواهد موجود نشان می دهد در آن زمان ریاضیدانان مدل‌های آماری تهیه می‌کردند و در اختیار سرداران نظامی قرار می دادند تا برای تصمیم‌گیری در میدان‌های جنگ استفاده گردد[1].

به نظر می رسد در سده اخیر برای اولین بار اصطلاح علم داده توسط پیتر نائور[1] مطرح شد. وی در سال ۱۹۶۰میلادی علم داده را به جای علم کامپیوتر به کار برد. نائور بعدها اصطلاح "داده شناسی"[2] را بدین منظور معرفی کرد و در سال ۱۹۷۴ در مقاله‌ای با عنوان "بررسی دقیق روش‌های کامپیوتری" از اصطلاح علم داده برای بیان پردازش‌های داده که در آن دوران در زمینه های وسیعی کاربرد داشت، استفاده کرد[2].

در سال ۱۹۹۶، اعضای "فدراسیون بین‌المللی جوامع دسته‌بندی"[3] در گردهمایی دو سالانه خود، در شهر کوبه ژاپن گرد هم آمدند. در این گردهمائی اصطلاح علم داده توسط چیوی هوایشی[4] معرفی شد و از آن پس علم داده به‌عنوان "علم داده، دسته‌بندی و روش‌های مرتبط" استفاده گردید[3].

در نوامبر سال ۱۹۹۷، سی اف جف وو[5] در افتتاحیه همایشی در حوزه علم داده، سخنرانی خود را با عنوان "آمار یا علم داده" در دانشگاه میشیگان ارائه داد[4]. وی در این سخنرانی از کارهای آماری با عنوان سه‌گانه ی "گردآوری، مدل‌سازی و تحلیل داده" یادکرد و میان علم داده و دانش آمار پیوند و نزدیکی برقرار کرد. در

---

[1] Peter Naur

[2] datalogy

[3] International Federation of Classification Societies | IFCS

[4] Chikio Hayashi

[5] C. F. Jeff Wu

این راستا، جف وو نام آمار را به علم داده و آماردان را به دانشمندان داده تغییر داد.

ویلیام کلیولند[1] در سال ۲۰۰۱ در مقاله خود به نام "علم داده" برنامه‌ای برای گسترش جنبه‌های فنی در رشته آمار پیشنهاد کرد تا علم داده به‌عنوان یک رشته مستقل شناخته شود. کلیولند این رشته جدید را مرتبط با علم کامپیوتر و داده‌کاوی می‌دانست[5]. وی بر این باور بود که منافع استفاده از یک تحلیلگر داده محدود است؛ زیرا از یک سو مهندسین کامپیوتر شناخت محدودی نسبت به روش‌های کار با داده دارند و از سوی دیگر دانش محاسباتی متخصصین آمار محدود است. بنابراین تلفیق این دو گروه می‌تواند منجر به نوآوری‌های بیشتری در زمینه علم داده گردد. برای تحقق این مقصود، کلیولند پیشنهاد داد تا اساتیدی در دپارتمان‌های علم داده فعالیت کنند که قادر باشند دانش داده را با دانش محاسبات تلفیق کنند.

در سال ۲۰۰۷ "جیم گری"[2] برنده جایزه تورینگ، علم داده را به‌عنوان چهارمین پارادایم علم معرفی کرد که از تحلیل نتایج اجرای محاسبات بر روی داده‌ی بزرگ حاصل می شود؛ در کاربرد اخیر علم داده روشی است که در آن می توان با استفاده از داده‌های موجود، دانش جدیدی را بنا نهاد و آمار جدیدی را تولید کرد[6].

دی جی پاتیل[3] در سال ۲۰۱۲ در مقاله "دانشمند داده: جذاب‌ترین شغل قرن ۲۱" که در مجله "کسب وکار" هاروارد منتشر شد، ادعا کرد که همراه با جف همر باکر[4] برای اولین بار این عبارت را در سال ۲۰۰۸ برای معرفی شغل خود در لینکدین و فیس‌بوک ابداع کرده‌اند. وی از دانشمندان داده به‌عنوان نسلی جدید یادکرد که فقدان آن‌ها منجر به محدودیت‌های جدی در بخش‌هائی از صنعت و دانشگاه می‌شود[7].

از دیگر نمونه های اهمیت و گسترش علم داده می‌توان به انجمن آمار آمریکا اشاره کرد که عنوان ژورنال خود را به "تحلیل‌های آماری و داده‌کاوی: " تغییر نام داد. مدتی بعد و طی تغییر نامی دوباره، نام بخش اول این ژورنال به "یادگیری آماری و علم داده" مبدل شد[8].

در سال ۲۰۱۵ چاپ " مجله بین‌المللی علم و تحلیل داده" توسط اسپرینگر[5] به‌منظور انتشار کارهای انجام شده در حوزه علم داده و تحلیل داده‌ی کلان آغاز شد و در سپتامبر همان سال طی سومین اجلاس ECDA در دانشگاه اسکس[6] عبارت «Gesellschaft für Klassifikation» به نام انجمن علم داده افزوده شد[9].

---

[1] William Klio Land

[2] Jim Gray

[3] Dhanurjay “DJ” Patil

[4] Jeff Hammerbacher

[5] springer

[6] Essex

۲-۳- تعریف علم داده

در زبان فار سی کلمه لاتین "science" به علم و "data" به داده ترجمه می شود. از ترکیب این دو کلمه به علم داده یاد می‌شود. علم داده[1] دانشی میان‌رشته‌ای است که متمرکز بر استخراج دانش و علم از مجموعه‌ی داده‌ی موجود اســت. علم داده در واقع دانشــی اســت که از ترکیب داده، تحلیل و اســتنتاج از آن به دســت می‌آید؛ یا به عبارت بهتر دانشی است که با ساختار معینی از اطلاعاتی که مبتنی بر آزمایش و واقعیت است و با استفاده از پیش‌بینی و الگوریتم به دست می‌آید[10]. علم داده از ترکیب مباحث مختلفی به وجود آمده و بر مبانی و روش‌های موجود در حوزه‌های مختلف علمی بنا شده است. تعدادی از این حوزه‌ها عبارت‌اند از: ریاضیات، آمار، مهندسی داده، باز شناخت الگو و ... . هدف این علم، استخراج مفاهیم جدید از داده و تولید "دانش جدیدِ داده محور" است.

علم داده به‌شدت در سال‌های اخیر رشد داشته است. نه تنها افزایش حجم داده، بلکه فرآیندهای موجود در علم داده، منجر به تولید دانش شده است و این چرخه تبدیل داده به دانش نیز خود به رشد مضاعف این علم کمک شایان توجهی کرده است.

علم داده در موضوعات بسیار گسترده‌ای کاربرد دارد؛ در فضانوردی، پزشکی و حتی زندگی روزمره استفاده می‌شود. بشر هر روزه از تکنولوژی استفاده می‌کند و تکنولوژی و سرویس‌های مبتنی بر آن می‌تواند اطلاعات کاربری افراد را جمع‌آوری‌کند.

گسترده‌تر شدن داده فقط در مقدار و حجم داده نبوده، بلکه تعداد متغیرهای هر مجموعه داده نیز رشد یافته است و این امر سبب گسترش علم داده در ابعاد گوناگون با سرعت زیادی گردیده است. لازم به ذکر است که علاوه بر رشــد کمی داده از جنس عدد و رشــته، در ســال‌های اخیر داده از جنس‌های دیگر، مانند: محتوای اینترنتی، محتوای چندرسانه‌ای و ... نیز تولید و گسترش یافته است. در گزارش IBM آمده: «۷۵٪ داده‌ی تولیدشده، ساختار نیافته و منابعی مانند: متن، صدا و ویدئو هستند»[11]. بنابراین گسترش استفاده انسان‌ها از تکنولوژی و تولید محتوا از یک ســو و به کار گیری الگوریتم های آماری و گســترش محاســبات ریاضــی از سوی دیگر سبب رشد چشم گیر علم داده شده است.

[1] Data Science

۲-۴- متخصص علم داده[1]

به شاغلین در حوزه‌ی علم داده، متخصص علم داده یا دانشمند داده می‌گویند. دانشمند داده کسی است که دارای مهارت‌هایی مثل تحلیل داده، مدل‌سازی و آمار باشد. در کنار این مهارت‌ها، کار دیگری هم جزو وظایف اصلی یک متخصص داده است، که از آن به‌عنوان مرتب‌سازی داده یاد می‌شود. به‌طور خلاصه منظور از مرتب‌سازی داده این است که متخصص این دانش بتواند داده‌ی به‌هم‌ریخته را ابتدا مرتب و منظم کرده و سپس به بررسی آن‌ها بپردازد[12] که در ادامه تک‌تک این وظایف را به‌صورت موردی بررسی کند.

چن فو جف وو در سال ۱۹۹۸ برای اولین بار در یک سخنرانی از واژه‌ی متخصص علم داده استفاده کرد. متخصصین علم داده با تعمق در چندین رشته‌ی علمی، مسائل پیچیده‌ی مطرح‌شده در حوزه‌ی داده را حل می‌کنند. به‌طورکلی انتظار می‌رود که متخصصین علم داده قادر باشند در بخش‌هایی از علم ریاضیات، آمار و کامپیوتر کار کنند. یک متخصص علم داده می‌بایست در یک یا دو رشته تخصص داشته باشد و در دیگر حوزه‌ها دارای مهارت کافی باشد.

برخی معتقدند متخصص علم داده توسط دی جی پاتیل و جف همرباکر ابداع شده است. دی جی پاتیل در سال ۲۰۱۲ در مقاله "علم داده: جذاب‌ترین شغل قرن بیست و یکم" متخصصین علم داده را این‌طور تعریف می‌کند: کسانی که می‌دانند چگونه می‌توان از انبوه اطلاعات بدون ساختار، پاسخ سؤال‌های کسب وکار را پیدا کرد.

استنتون در سال ۲۰۱۳ در تعریف علم داده به مهارت های لازم برای متخصص بودن در این علم اشاره می کند: علم داده رشته‌ی در حال ظهوری است که به جمع‌آوری، آماده‌سازی، تحلیل، بصری سازی، مدیریت و نگهداشت اطلاعات در حجم بالا می‌پردازد.

دریسکول در سال ۲۰۱۴ در تعریف علم داده و متخصص آن می نویسد: علم داده مهندسی ساختمان داده است. و متخصص علم داده دانشی کاربردی از داده و ابزارها دارد. به‌علاوه درک تئوری دارد که مشخص می‌کند چه چیزی ازنظر علمی ممکن است[13].

بدین ترتیب هرکسی که به یافتن موقعیت شغلی قدرتمندتر در این حوزه علاقه‌مند است، باید مهارت‌های کلیدی در سه زمینه تحلیل، برنامه‌نویسی و دانش مربوط به حوزه مورد استفاده را کسب کند. با نگاهی عمیق‌تر، می‌توان گفت مهارت‌های زیر می‌تواند به افراد جهت تبدیل‌شدن به یک دانشمند داده کمک کند:

[1] Data Scientist