

بسمه تعالی

فرم ثبت نام در چالش  
"داده کاوی هوشمند به توان شناختی"

سال 1398

فرم ثبت نام در چالش: داده کاوی هوشمند به توان شناختی

## ۱ معرفی تیم شرکت کننده در چالش

ردیف	نام و نام خانوادگی	نقش در تیم	تخصص در تیم	رشته تحصیلی	آخرین مدرک تحصیلی	دانشگاه	موبایل	ایمیل
۱	سمیه گل محمدی	مدیر تیم	اقتصادسنجی و برنامه نویسی	علوم کامپیوتر	کارشناسی ارشد	علامه طباطبایی	۰۹۳۵۸۳۰۹۱۲۲	<a href="mailto:So_gmo@yahoo.com">So_gmo@yahoo.com</a>
۲	محمدامین خدامرادی	عضو	داده کاوی و برنامه نویسی	علوم کامپیوتر	کارشناسی ارشد	شهید بهشتی	۰۹۳۶۱۶۴۳۴۶۵	<a href="mailto:Khodamoradi1992@gmail.com">Khodamoradi1992@gmail.com</a>

اطلاعات کسب و کار نوپا یا شرکت دانش بنیان در صورت عضویت تیم:

نام کسب و کار	
آدرس وب سایت کسب و کار	
سال تشکیل کسب و کار	
نوآوری/محصول/خدمتی که کسب و کار ارائه می دهد	

## ۲ توصیف اجمالی طرح

### الف-مرور پژوهش (حداکثر 2000 کلمه):

از آنجایی که نمایش درست داده ها همیشه بخش زیادی از اطلاعات پنهان در مجموعه داده را به ما نشان می دهد. در نتیجه یکی از بخش های مهم مورد توجه ما در این پروژه نمایش اطلاعات مجموعه داده به شیوه های مختلف و سعی در پیدا کردن الگوهای مفید در بین آنهاست. برای این منظور ما مدنظر داریم مجموعه داده را به شکل های زیر نمایش دهیم:

#### فاز اول ( بدون عامل زمان یا پولینگ روی روز و ساعت)

##### مرحله اول

- 1- نمایش نقشه جغرافیایی مکان هایی که راننده درخواست را پذیرفته است و سفر بصورت کامل انجام شده است.
  - 2- نمایش نقشه جغرافیایی مقاصدی که راننده درخواست را پذیرفته و سفر بصورت کامل انجام شده است.
  - 3- نمایش نقشه جغرافیایی مکان هایی که راننده درخواست را پذیرفته است ولی سفر انجام نشده است.
  - 4- نمایش نقشه جغرافیایی مقاصدی که راننده درخواست را پذیرفته ولی سفر انجام نشده است.
  - 5- نقشه جغرافیایی مکان هایی که برای آنها راننده ای پیدا نشده است.
  - 6- نقشه جغرافیایی مقاصدی که برای آنها راننده ای پیدا نشده است.
- در مرحله اول تمام این نقشه ها بدون در نظر گرفتن زمان رسم شده و به بررسی الگوهای معناداری می پردازیم که میان آنها وجود دارد.

##### مرحله دوم

- رسم دوتا از متغیر ها بصورت همزمان به این صورت که این بار با یک سری بردار که جهت شروع و پایان آن هرکدام رنگ مجزایی دارد ،
- نقشه سفرهایی که بصورت کامل انجام شده است را به صورت بردار جهت داری رسم و به دنبال الگویی می گردیم که نشان دهد کدام مبدا به مقصدا بصورت مکرر تکرار شده است.
  - نقشه مبدا و مقصد سفرهایی که انجام نشده است را به صورت بردار جهت داری رسم و به دنبال الگویی می گردیم که نشان دهد کدام مبدا به مقصدا بصورت مکرر تکرار شده است .
  - نقشه مبدا و مقصد سفرهایی که راننده ای آنها را نپذیرفت را به صورت بردار جهت داری رسم و به دنبال الگویی می گردیم که نشان دهد کدام مبدا به مقصدا بصورت مکرر تکرار شده است .

## فرم ثبت نام در چالش: داده کاوی هوشمند به توان شناختی

### فاز دوم (با در نظر گرفتن عامل زمان و پولینگ روی روزهای هفته)

در مرحله بعدی با خوشه بندی ساعت ها بر اساس تعداد تراکنش سه دسته ی ساعات پرتراکنش، معمولی، ساعات کم تراکنش را پیدا و متغیر زمان را به این سه دسته تقسیم می کنیم. و تمام مراحل اول و دوم را مجددا تکرار و الگوهای موجود را استخراج نماییم.

### فاز سوم (با در نظر گرفتن روزهای هفته)

این فاز که جامع ترین حالت را در نظر می گیرد . تعداد روزهای هفته را لحاظ می کنیم(ممکن است باتوجه به زیاد بودن مجموعه داده ،مجموعه داده برای افزایش عملکرد و کارایی مدل مجبور به تقسیم روزهای هفته به دو دسته روزهای عادی و آخر هفته بشویم. ولی ترجیحا هر روز به صورت جداگانه بررسی می شود). ساعت تراکنش را هم به سه خوشه تقسیم کرده (به صورتی که در فاز دوم مطرح شد) در نظر می گیریم. به این ترتیب نمودار سه بعدی خواهیم داشت که یک بعد روزهای هفته و یک بعد زمان و بعد سوم یک ویژگی دلخواه خواهد بود.

1-ابتدا ماتریس تعداد تراکنش را در روز و ساعت های مختلف هفته می توان رسم کرد.نقشه می تواند دوبعدی و متغیر سوم با رنگ مشخص میشود. پررنگ بود نشانه ی وجود تعداد بیشتری تراکنش در آن زمان و روز مشخص است.

2- در این مرحله متغیر سوم یکی از شش حالت زیر می تواند باشد لذا به بررسی الگوهای ممکن در ساعت و روزهای مختلف می پردازیم.

- نمایش نقشه جغرافیایی مکان هایی که راننده درخواست را پذیرفته است و سفر بصورت کامل انجام شده است.

- نمایش نقشه جغرافیایی مقاصدی که راننده درخواست را پذیرفته و سفر بصورت کامل انجام شده است.

- نمایش نقشه جغرافیایی مکان هایی که راننده درخواست را پذیرفته است ولی سفر انجام نشده است.

- نمایش نقشه جغرافیایی مقاصدی که راننده درخواست را پذیرفته ولی سفر انجام نشده است.

- نقشه جغرافیایی مکان هایی که برای آنها راننده ای پیدا نشده است.

- نقشه جغرافیایی مقاصدی که برای آنها راننده ای پیدا نشده است.

### فاز چهارم ( اطلاعات متفرقه )

در این مرحله با توجه به منطق مجموعه داده به دنبال کشف الگوهایی هستیم که احتمالا در رفتار راننده یا کاربر تاثیر گذار هستند.

- نمایش همبستگی بین متغیر isAccept و متغیر driverETA

- نمایش همبستگی بین متغیر rideUpdatedAt و driverETA

چرا که به احتمال زیاد این متغیرها به شدت به هم همبسته باشند.

## فرم ثبت نام در چالش: داده کاوی هوشمند به توان شناختی

### فاز پنجم) تاثیر سلیقه و رفتار راننده)

استفاده از بار چارت

نمایش تعداد تراکنش های پیشنهادی به یک راننده مخصوص .

(بدون در نظر گرفتن ساعت و روز، با در نظر گرفتن فقط ساعت، با در نظر گرفتن ساعت و روز )

- نمایش تعداد تراکنش هایی که راننده پذیرفته است.

- نمایش تعداد تراکنش هایی که راننده پذیرفته ولی سفر در نهایت انجام نشده است.

با نمایش این مقدارها می توان دسته بندی ویژه ای برای راننده ها در نظر گرفت به این صورت که راننده (کوشا- معمولی-ضعیف) عمل میکند. با توجه به اطلاعات موجود در مجموعه داده threshold مناسب و روش مناسب خوشه بندی اتخاذ می گردد.

این کار فواید زیادی به همراه دارد:

- 1- می توانیم قبل از درخواست به راننده تخمین بزنیم که آیا می پذیرد یا خیر، و درخواست را فقط برای رانندگانی ارسال کنیم که احتمال پذیرش بالاتری دارند. در نتیجه در مدت زمانی که کاربر منتظر می ماند تا درخواستش قبول شود صرفه جویی می شود. چرا که اگر هر راننده زمان مثلا 30 ثانیه به این موضوع بیندیشد که آیا سفر را تایید و لغو کند، در برای حالتهایی که لازم است درخواست به چندین راننده ارسال شود، صرفه جویی زمانی برای مسافر و صرفه جویی تصمیمی برای راننده خواهیم داشت.
- 2- با این مکانیسم به رانندگانی که کوشا تر هستند عملا پاداش داده می شود چرا که احتمال پذیرش بالاتری برای درخواست دارند.

نکات مهم:

برای اجرای این حالت ابتدا بستر لازم باید فراهم شود به این صورت که الگوریتم درخواست را برای نزدیک ترین راننده ارسال نکند، بلکه دسته ای از راننده های بالقوه را در نظر گرفته و آنها را امتیاز بندی کند و طبیعتا یکی از مهم ترین ویژگی های این دسته این است که فاصله زمانی کمترین باشد، ولی ویژگی های دیگری چون سابقه خوب راننده هم دخیل خواهد بود. یکی دیگر از ویژگی هایی که میتواند لحاظ کرد ، زمان آخرین تراکنش قبلی است. به این صورت که ویژگی تراکنش- ای دی راننده را تشکیل و از طریق آن متوجه شویم آیا حضور راننده در این قسمت و لوکیشن بخاطر سفر قبلی تبسی بوده است یا خیر. این ویژگی اگر مثبت باشد امتیاز بالاتری به راننده برای انتخاب شدن جهت ارسال درخواست می دهد.

فایده اضافه کردن این ویژگی این است که رانندگان به راحتی مقاصدی مانند فرودگاه ... را قبول می کنند که چرا که در می یابند اگر بخاطر سفر تبسی به این مقصد بروند احتمال برگشت خالی شان کمتر از قبل است. و شانس بالاتری نسبت به رانندگان از تبسی دارد که به دلایل شخصی در آن منطقه حضور یافته اند.

## فرم ثبت نام در چالش: داده کاوی هوشمند به توان شناختی

### ب- تشریح ایده (حداکثر 2000 کلمه):

در این پژوهش برآنیم از مجموعه داده‌ای که توسط شرکت Tap30 برای این چالش فراهم شده استفاده نماییم. بعد از مطالعه و بررسی داده‌ها به این نتیجه رسیدیم که می‌توان تحلیل‌هایی بصری از رفتار راننده‌ها و مسافران برای این مجموعه داده ارائه داد. با به کارگیری ابزارهای مصورسازی و زبان‌های برنامه‌نویسی هم‌چون R و Python در صدد خواهیم بود این تحلیل‌ها را فراهم و ارائه دهیم. تحلیل‌ها به دو دسته‌ی مبتنی بر زمان و بدون درنظر گرفتن زمان تقسیم می‌شود. این تحلیل‌ها در جهت پاسخگویی به سؤالاتی که در ادامه ذکر خواهد شد ارائه می‌شوند:

- 1 نحوه‌ی توزیع مکان اولیه‌ی مسافران برروی نقشه چگونه است؟
- 2 نحوه‌ی توزیع مقاصد مسافران برروی نقشه چگونه است؟
- 3 توزیع انواع سفرهای کنسل شده، راننده یافت نشد و به اتمام رسیده روی نقشه چگونه است؟ آیا رابطه‌ی معناداری بین موقعیت جغرافیایی و وضعیت نهایی سفر وجود دارد؟
- 4 نمودار حرارتی درخواست سفر در روزهای هفته و ساعت‌های روز به چه شکل بوده؟ و زمان‌های کم درخواست و پر درخواست در بازه‌ی یک هفته چه زمان‌هایی هستند؟
- 5 اگر داده‌ی مربوط به وضعیت آب و هوا را داشته باشیم، رفتار تقاضای مسافران و نحوه‌ی درخواست دادن آن‌ها در وضعیت‌های مختلف آب‌وهوایی چگونه خواهد بود؟ و چه مقدار تغییر خواهد کرد؟
- 6 رفتار مسافرانی که سفر آن‌ها توسط راننده قبول شده چگونه است؟ چه مسافرانی، سفر خود را لغو می‌کنند؟ یک مدل یادگیری ماشین می‌تواند این رفتار را تحلیل کرده و پیش بینی نماید.
- 7 اگر داده‌های سفرهای پیشنهادی به هر راننده را داشته باشیم، آیا می‌توان پیش بینی کرد کدام راننده یک درخواست سفر را قبول می‌کند؟ یک مدل یادگیری ماشین می‌تواند این رفتار را تحلیل و پیش بینی نماید.
- 8 اگر اطلاعات زمان پایان آخرین سفر قبلی راننده را داشته باشیم، آیا زمان انتظار راننده به عنوان یک المان و ویژگی می‌تواند در مدل یادگیری ماشین درنظر گرفته شده و مدل را تقویت کند؟

پاسخ‌گویی به سؤالات مذکور و ارائه نمودارها و شکل‌های خواسته شده می‌تواند نقش مهمی در تحلیل رفتاری مسافران و رانندگان داشته باشد. به گونه‌ای که برای هر روز هفته و ساعت‌های روز با افزایش و یا کاهش موقتی قیمت‌ها باعث جذب و یا دفع راننده‌ها شود. یا با استفاده از پیام‌های نرم‌افزار، راننده‌هایی که زمان زیادی منتظر بوده‌اند را از مکان‌هایی که تقاضا بیش از عرضه است مطلع کند.

همچنین می‌توان سیستم پیشنهاد سفر به راننده را تغییر داده و با درنظر گرفتن المان‌های بیشتر نسبت به قبل تقویت نمود.

### فرم ثبت نام در چالش: داده کاوی هوشمند به توان شناختی

داده‌های زیر همان‌طور که در متن شرح داده شد، در صورت وجود و در اختیار بودن می‌توانند تحلیل‌ها را تقویت کرده و در ارائه‌ی مدل رفتاری بهتر از مسافران و رانندگان کمک کنند:

- 1 داده‌ی مربوط به وضعیت آب و هوا در زمان درخواست سفر
- 2 داده‌های سفرهای پیشنهادی به هر راننده
- 3 اطلاعات زمان پایان آخرین سفر قبلی راننده (زمان انتظار راننده برای سفر بعدی)

#### ج- تشریح کیفیت و مراحل اجرای طرح (حداکثر 1000 کلمه):

ابتدا لازم است، داده‌ها به طور کامل از شرکت Tap30 دریافت شده و درمورد سیستم محاسبه‌ی driverETA اطلاعات کافی جمع‌آوری شود. سپس داده‌هایی که مقدار آن‌های گذاشته شده یا مقادیر غیرمتعارف دارند سازماندهی و در صورت لزوم حذف شوند. داده‌های به شکلی آماده شوند که بتوان نمودارها و شکل‌های مذکور را از روی آن‌ها رسم کرده و ارائه داد. سپس تحلیل‌های هر شکل را بیان کرده و این تحلیل‌های براساس روابط آماری اعتبارسنجی شده و معنی‌دار می‌شوند. در مرحله‌ی بعدی مدل‌های یادگیری ماشین مختلف اعمال شده و مدلی با دقت و توان پیش‌بینی بالاتر بر روی داده‌های آزمون معرفی می‌شود و با بقیه‌ی مدل‌ها مقایسه خواهد شد. جهت سنجش مقدار قابل اعتماد بودن مدل و ثبات آن، مدل در یک ساز و کار اعتبار سنجی متقابل (kfold\_CV) بر روی داده‌های آزمون، آزموده می‌شود. معیارهایی هم‌چون AUC، AUPR و Fmesear ملاک‌های ارزیابی مدل یادگیری ماشین خواهند بود.

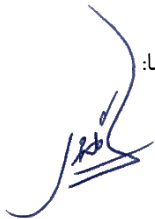
#### د- دادگان به کار گرفته شده در این طرح :

مجموعه داده‌ی Tap30

## ۳ تعهدات

اینجانب سمیه گلمحمدی به عنوان مدیر تیم با امضای این فرم کلیه مندرجات آن‌را از نظر صحت و دقت تایید می‌نمایم.

فرم ثبت نام در چالش: داده کاوی هوشمند به توان شناختی

نام و نام خانوادگی: سمیه گل محمدی	امضا: 	تاریخ: ۱۳۹۸/۰۸/۲۴
-----------------------------------	---	-------------------