WIFI LATENCY ANALYSIS REPORT

Aryan Nourbakhsh

BNUT | BABOL NOSHIRVANI UNIVERSITY OF TECHNOLOGY
2025 Spring

1. مقدمه

هدف این پروژه تحلیل داده‌های مربوط به اتصال کاربران به شبکه‌های وای‌فای و بررسی عواملی است که می‌توانند بر مقدار تأخیر (latency) در این شبکه‌ها تأثیرگذار باشند. تأخیر به عنوان یکی از معیارهای اصلی کیفیت سرویس در شبکه‌های بی‌سیم شناخته می‌شود و نقش مهمی در تجربه کاربری به‌ویژه در کاربردهای حساس به زمان مانند تماس تصویری، بازی آنلاین و استریم دارد. در این پروژه سعی شده است با استفاده از تحلیل آماری، مصورسازی داده‌ها و محاسبات اطلاعاتی، متغیرهای تأثیرگذار بر تأخیر شناسایی و میزان اهمیت آن‌ها تعیین شود.

2. معرفی داده‌ها

داده‌های این پروژه در قالب فایل csv شامل نمونه‌هایی از اتصال کاربران به شبکه‌های وای‌فای در شرایط مختلف می‌باشد. هر ردیف از داده‌ها مربوط به یک اتصال خاص است و ویژگی‌های مختلفی از آن اتصال را ثبت کرده است.

مجموعه داده شامل ویژگی‌های زیر است:

- latency_ms: مقدار تأخیر اتصال به میلی‌ثانیه
- rssi_dbm: قدرت سیگنال دریافتی به دسی‌بل میلی‌وات
- snr_db: نسبت سیگنال به نویز به دسی‌بل
- channel_util%: درصد استفاده از کانال بی‌سیم
- num_assoc_devices: تعداد دستگاه‌های متصل به نقطه دسترسی
- client_speed_mbps: سرعت اتصال کلاینت برحسب مگابیت بر ثانیه
- distance_m: فاصله بین دستگاه کاربر و نقطه دسترسی به متر
- band: باند فرکانسی استفاده‌شده 2.4GHz) یا 5(GHz
- Protocol: پروتکل وای‌فای مورد استفاده) مانند 802.11 nیا 802.11(ac
- ap_vendor: نام تولیدکننده سخت‌افزار نقطه دسترسی

3. پیش‌پردازش داده‌ها

در این بخش، داده‌ها برای تحلیل آماری و مدل‌سازی آماده‌سازی شدند. مراحل به ترتیب زیر انجام شدند:

3.1. انتخاب ویژگی‌ها

از بین تمامی ستون‌های موجود، تنها ستون‌هایی که به صورت مستقیم بر تأخیر شبکه تأثیرگذار هستند انتخاب شدند تا تمرکز تحلیل روی آن‌ها باشد.

3.2. حذف داده‌های ناقص

تمامی ردیف‌هایی که دارای مقادیر گمشده (NaN) در هر یک از ستون‌های مهم بودند حذف شدند تا از بروز خطا در مراحل بعدی جلوگیری شود.

3.3. حذف نقاط پرت آماری

برای تمامی ویژگی‌های عددی، از روش IQR (Interquartile Range) برای شناسایی و حذف مقادیر پرت استفاده شد. در این روش، بازه‌ای بین $Q1 - 1.5IQR$ و $Q3 + 1.5IQR$ به عنوان بازه مجاز تعریف می‌شود و مقادیری که خارج از این بازه باشند به عنوان پرت حذف می‌شوند.

فکر کنم متوجه نشده باشید. بذارید یه توضیح مفصل بهتر بدم.

در تحلیل داده‌ها، نقاط پرت به مقادیری گفته می‌شود که به طور غیرعادی با سایر مقادیر تفاوت دارند. این مقادیر می‌توانند ناشی از خطاهای اندازه‌گیری، شرایط غیرمعمول یا رویدادهای نادر باشند و در صورت عدم حذف، ممکن است باعث تحریف نتایج تحلیل‌های آماری یا مدل‌های پیش‌بینی شوند.

برای شناسایی و حذف این نقاط از روش IQR استفاده می‌شود .IQR یا "فاصله بین چارک‌ها"، تفاوت بین چارک اول (Q1) و چارک سوم (Q3) داده‌ها است. با محاسبه این فاصله، یک بازه قابل‌قبول برای داده‌ها تعیین می‌شود که حدود آن برابر است با:

- حد پایین = $Q1 - 1.5 \times IQR$
- حد بالا = $Q3 + 1.5 \times IQR$

مقدارهایی که خارج از این بازه قرار می‌گیرند، به عنوان داده پرت در نظر گرفته شده و از مجموعه داده حذف می‌شوند.

استفاده از روش IQR در مقایسه با روش‌هایی مانند استفاده از میانگین و انحراف معیار، به دلیل عدم حساسیت به توزیع داده و مقاومت در برابر چولگی، روشی مطمئن و مؤثر محسوب می‌شود. به ویژه در داده‌هایی که دارای مقادیر شدید یا غیرنرمال هستند، IQR می‌تواند نقش مهمی در پاک‌سازی داده و بهبود کیفیت تحلیل داشته باشد.

3.4. اعمال فیلترهای منطقی

برخی فیلترها براساس منطق و ماهیت فیزیکی ویژگی‌ها اعمال شد. از جمله:

- فاصله نمی‌تواند منفی باشد
- سرعت اتصال باید بزرگ‌تر یا مساوی صفر باشد
- درصد استفاده از کانال باید در بازه ۰ تا ۱۰۰ باشد
- نسبت سیگنال به نویز باید غیرمنفی باشد

3.5. تبدیل داده‌های دسته‌ای

برای بهینه‌سازی مصرف حافظه و افزایش کارایی پردازش، ستون‌های متنی مانندband ، protocolو ap_vendor به نوع داده category تبدیل شدند.

4. تحلیل ویژگی‌های دسته‌ای

با استفاده از نمودارهای جعبه‌ای(boxplot) ، توزیع تأخیر برای هر مقدار از سه ویژگی دسته‌ای یعنیband ، protocolو ap_vendor مورد بررسی قرار گرفت. این نمودارها نشان دادند که برخی دسته‌ها مانند پروتکل 802.11 acتأخیر کمتری نسبت به دسته‌های قدیمی‌تر دارند. همچنین تفاوت‌هایی بین تولیدکنندگان مختلف نقطه دسترسی مشاهده شد.

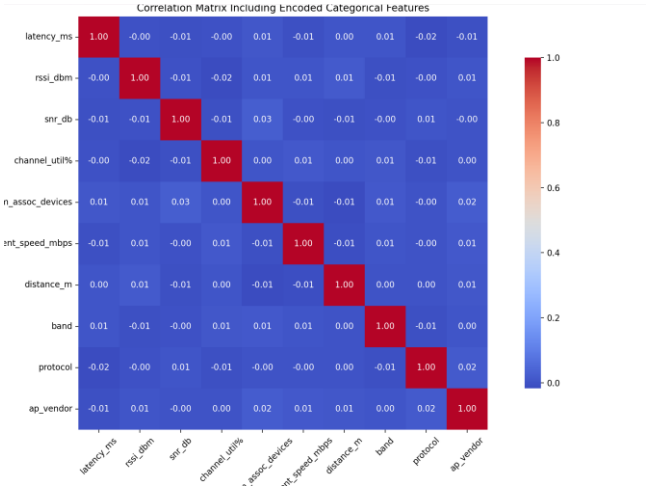
5. تحلیل ویژگی‌های عددی

برای تحلیل روند تأثیر ویژگی‌های عددی مانند فاصله، قدرت سیگنال و استفاده از کانال، هر ویژگی به ۱۰ بازه مساوی تقسیم شد و میانگین تأخیر در هر بازه محاسبه و به صورت نمودار خطی ترسیم شد. نتایج نشان داد که:

- با افزایش فاصله، میانگین تأخیر افزایش می‌یابد.
- با کاهش قدرت سیگنال(RSSI) ، تأخیر بیشتر می‌شود.
- نسبت سیگنال به نویز (SNR) با تأخیر رابطه معکوس دارد.

6. ماتریس همبستگی

برای بررسی همبستگی خطی بین ویژگی‌ها و تأخیر، ابتدا ویژگی‌های دسته‌ای به عدد تبدیل شدند) با استفاده از .(Label Encoding سپس ماتریس همبستگی (correlation matrix) محاسبه و به صورت heatmap نمایش داده شد. این تحلیل به طور کلی نشان‌دهنده روابط خطی بین ویژگی‌ها بود ولی نمی‌تواند رابطه‌های غیرخطی را نشان دهد.



7. توزیع لگاریتمی تأخیر

از آن‌جا که توزیع تأخیر دارای چولگی زیاد است(skewed distribution) ، از تبدیل لگاریتمی با تابع log1p برای نرمال‌سازی توزیع استفاده شد. سپس هیستوگرام این داده‌ها ترسیم شد تا توزیع بهتر قابل مشاهده باشد.

8. محاسبه اطلاعات متقابل(Mutual Information)

برای بررسی وابستگی کلی بین ویژگی‌ها و تأخیر (چه خطی چه غیرخطی)، از روش اطلاعات متقابل استفاده شد. مراحل به شرح زیر است:

8.1. باین‌بندی داده‌ها

تمامی ویژگی‌های عددی با استفاده از KBinsDiscretizer به ۶ بازه مساوی تقسیم شدند. همچنین تأخیر لگاریتمی به ۶ بازه به صورت صدکی (quantile-based) تقسیم شد تا کلاس هدف برای تحلیل اطلاعات متقابل آماده شود. بذارید بیشتر توضیح بدم KBinsDiscretizer یک ابزار از کتابخانه Scikit-learn است که داده‌های عددی پیوسته را به بازه‌های گسسته (بندی‌شده) تبدیل می‌کند. این ابزار به‌ویژه زمانی مفید است که بخواهیم ویژگی‌های پیوسته را برای تحلیل‌های طبقه‌بندی یا محاسبه اطلاعات متقابل (Mutual Information) به شکل دسته‌بندی‌شده درآوریم.

در این روش، مقدار هر ویژگی عددی به یکی از چند باکس یا بازه مشخص اختصاص داده می‌شود. تعداد بازه‌ها توسط کاربر تعیین می‌شود (مثلاً ۶ بازه)، و می‌توان نحوه تقسیم‌بندی را به سه شکل مختلف انجام داد:

- **Uniform**: تقسیم یکنواخت بازه‌ها بر اساس فاصله عددی
- **Quantile**: تقسیم بر اساس تعداد نمونه‌ها در هر بازه (هم‌اندازه از نظر تعداد)
- **Kmeans**: تقسیم با استفاده از خوشه‌بندیK-Means

در این پروژه از روش uniform استفاده شده که مقدار کل بازه به بازه‌های مساوی بر اساس فاصله عددی تقسیم می‌شود.

کاربرد اصلی آن در این پروژه، تبدیل ویژگی‌های عددی مانند فاصله، قدرت سیگنال و غیره به دسته‌های گسسته برای امکان مقایسه و محاسبه میزان وابستگی آن‌ها به تأخیر باین‌شده (هدف) بوده است. این کار کمک می‌کند که داده‌های پیوسته برای تحلیل‌های آماری یا اطلاعاتی آماده و قابل پردازش شوند.

8.2. رمزگذاری ویژگی‌های دسته‌ای

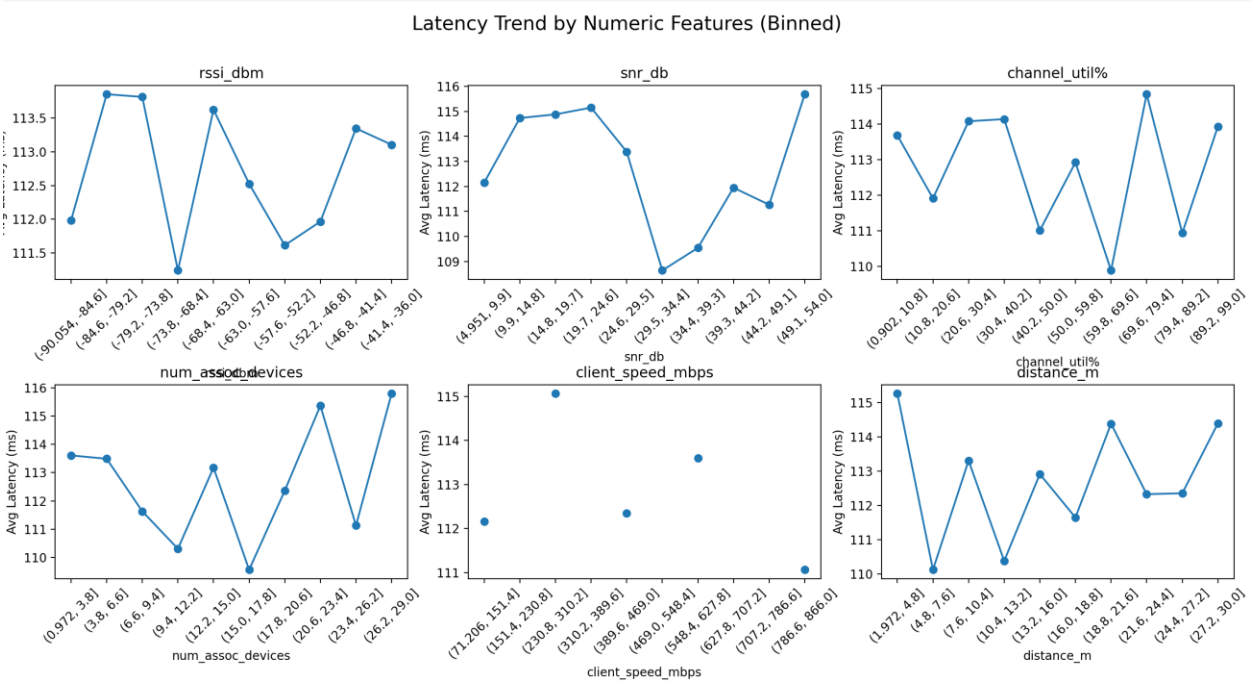
ویژگی‌های متنی به صورت عددی (Label Encoding) تبدیل شدند.

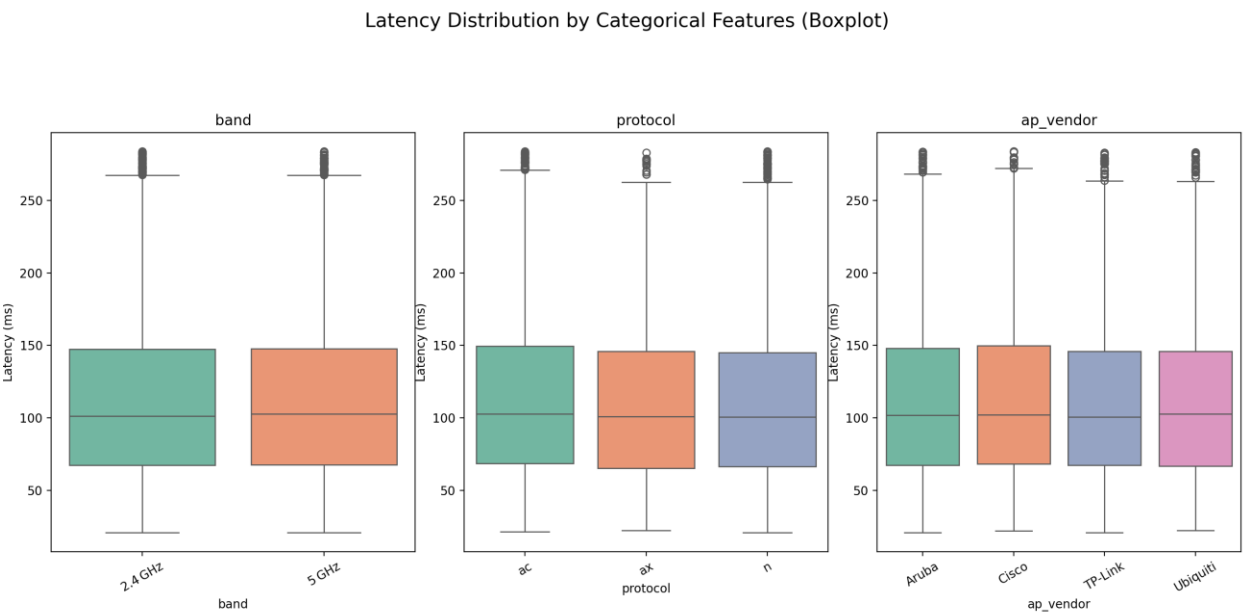
8.3. پیاده‌سازی تابع اطلاعات متقابل

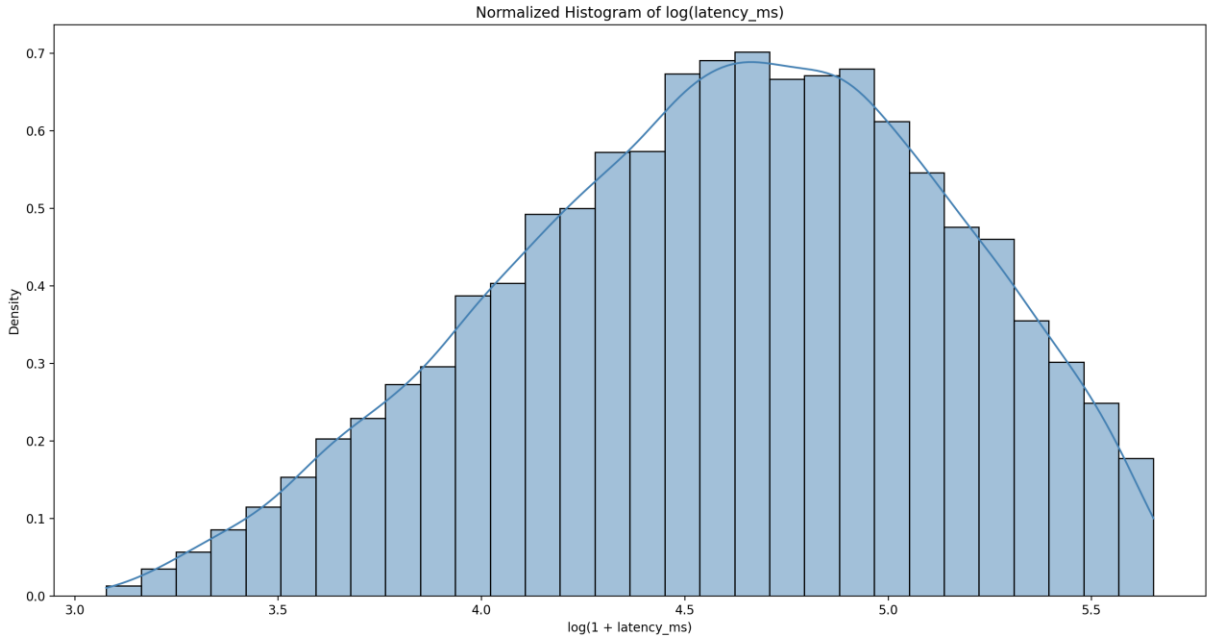
تابعی برای محاسبه اطلاعات متقابل بین دو متغیر به صورت دستی پیاده‌سازی شد تا بتوان دقیقاً نحوه محاسبه را کنترل کرد و وابستگی بین هر ویژگی و کلاس تأخیر را اندازه‌گیری نمود.

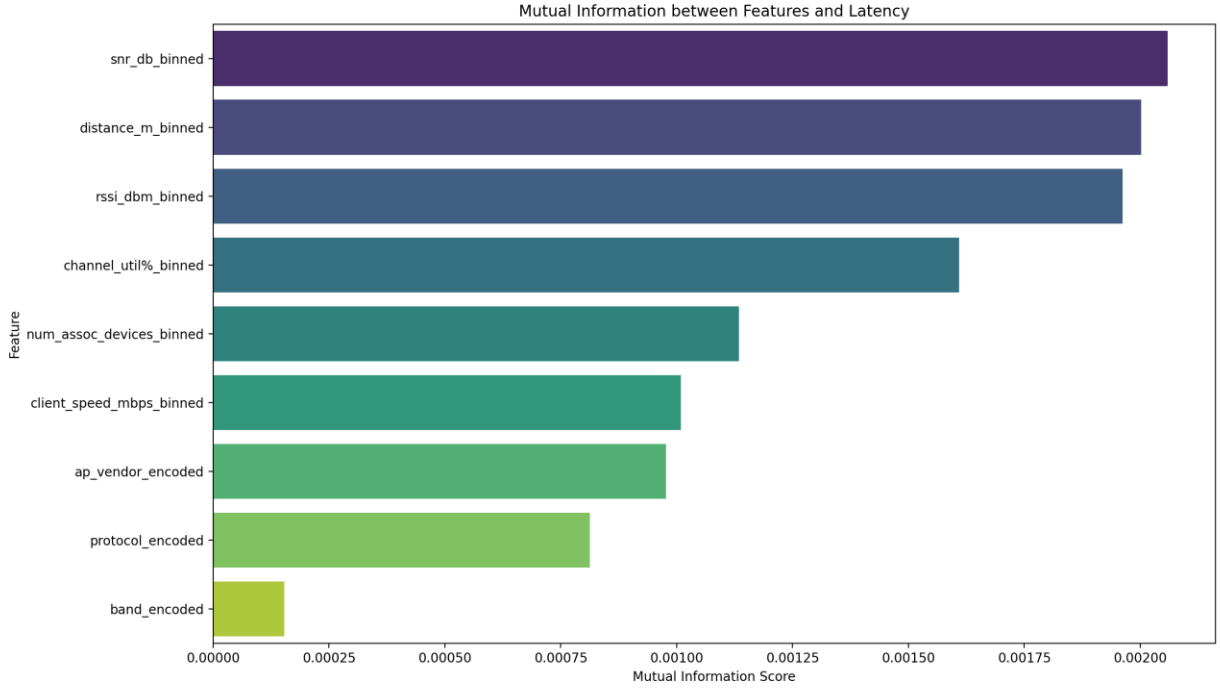
8.4. ترسیم Heatmap و نمودار میله‌ای

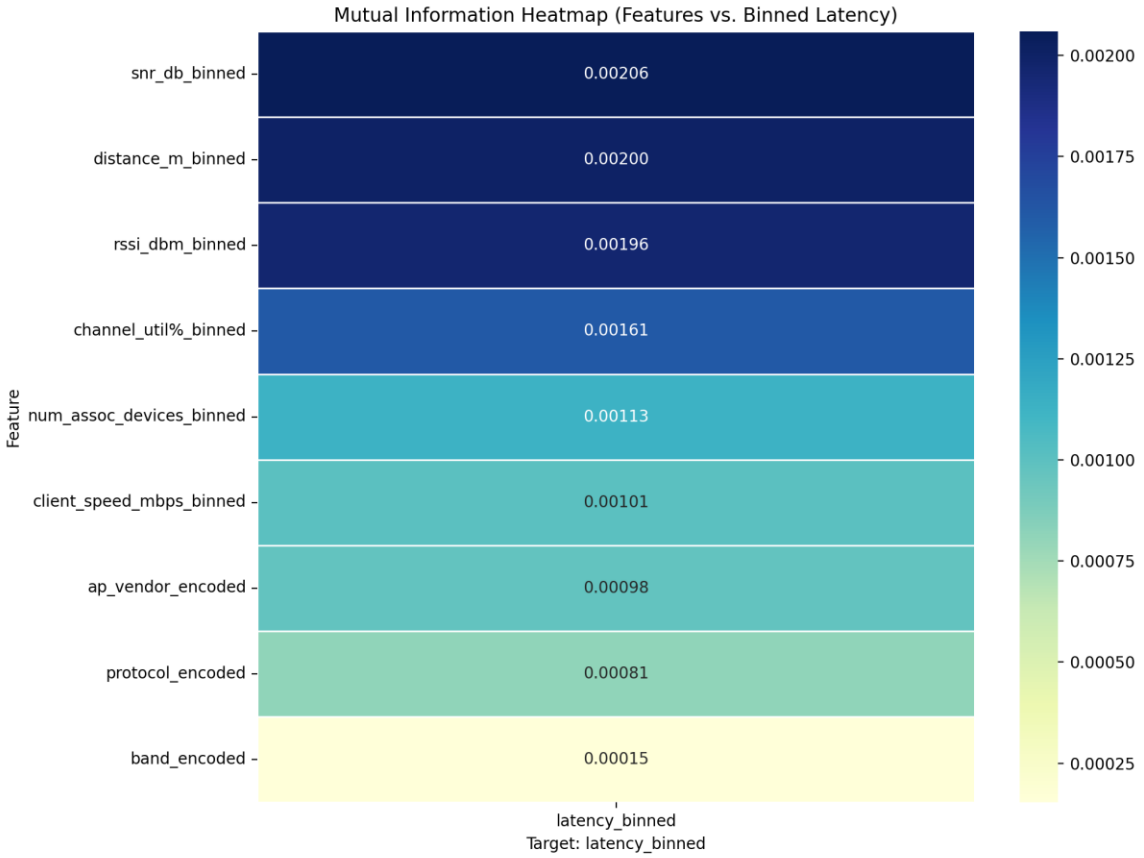
اطلاعات متقابل بین هر ویژگی و تأخیر به صورت heatmap و همچنین نمودار میله‌ای نمایش داده شد. نتایج نشان داد که ویژگی‌هایی مانند فاصله، SNRو RSSI بیشترین اطلاعات را در مورد تأخیر دارند.






Normalized Histogram of log(latency_ms)
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
3.0
3.5
4.0
4.5
5.0
5.5
log(1 + latency_ms)


Mutual Information between Features and Latency
snr_db_binned
distance_m_binned
rssi_dbm_binned
channel_util%_binned
num_assoc_devices_binned
client_speed_mbps_binned
ap_vendor_encoded
protocol_encoded
band_encoded
Feature
0.00000
0.00025
0.00050
0.00075
0.00100
0.00125
0.00150
0.00175
0.00200
Mutual Information Score



بر اساس تحلیل اطلاعات متقابل میان ویژگی‌ها و تأخیر شبکه، نتایج به شرح زیر است:

1. ویژگی‌های snr_db_binned، distance_m_binned و rssi_dbm_binned بیشترین میزان وابستگی به تأخیر دارند و به‌عنوان مهم‌ترین عوامل مؤثر شناسایی شدند. این ویژگی‌ها مستقیماً به کیفیت سیگنال و شرایط فیزیکی ارتباط مربوط می‌شوند.
2. ویژگی‌های channel_util%_binned، num_assoc_devices_binned و client_speed_mbps_binned وابستگی متوسطی به تأخیر دارند. این عوامل بیشتر به وضعیت محیطی و ترافیک شبکه مرتبط هستند.
3. ویژگی‌های ap_vendor_encoded، protocol_encoded و band_encoded کمترین میزان اطلاعات متقابل را با تأخیر دارند. تأثیر این ویژگی‌ها محدودتر و بیشتر غیرمستقیم است.

کد کامل پیاده‌سازی:

https://github.com/Aryanoor/Wifi-Latency-Analysis.git