
بسم الله الرحمن الرحيم

دانشگاه صنعتی شریف
دانشکده مهندسی صنایع

پروژه‌ی مدیریت فرآیندهای کسب‌وکار
رشته‌ی مهندسی صنایع، گرایش مدیریت مهندسی

تجزیه و تحلیل فرآیند درخواست وام بانکی

نگارش

نازنین قائمی‌زاده

استاد

جناب آقای دکتر عرفان حسن‌نایبی

پاییز ۱۴۰۲

فهرست مطالب

فصل اول: کلیات

۱- مقدمه

در سال‌های اخیر، افزایش چشمگیری در حجم اطلاعات رخ داده است. با توجه به این‌که قیمت دستگاه‌های ذخیره‌سازی در طول سال‌ها کاهش یافته است، ذخیره میلیون‌ها رکورد اطلاعات به یک امر رایج و مقرون به صرفه تبدیل شده است. ولیکن حجم زیاد داده، مشکلات جدی در استخراج اطلاعات ارزشمند ایجاد می‌کند و تجزیه و تحلیل مجموعه‌های داده به امر بسیار پیچیده‌ای مبدل گردیده است.

شرکت‌ها اغلب کنترل زیرفرآیندهایی که محصولات یا خدمات آن‌ها را تشکیل می‌دهند را ندارند. این موضوع در جریان‌های کاری با وجود وظایف تکراری منجر به افزایش هزینه‌ها شده و تأخیر در تحویل یک محصول یا خدمت نهایی به مشتری را موجب می‌شوند.

در این پروژه، هدف این است که بر مبنای یک *Event Log* واقعی متعلق به بانکی در هلند، فرآیند درخواست وام بانکی مدلسازی گردد. که به یک بانک نمونه در هلند تعلق دارد. مجموعه‌ی داده‌های موجود از رویدادها بسیار غنی است و در کل شامل ۲۶۲۲۰۰ رکورد (*Log*) ثبت شده است که به ۱۳۰۸۷ درخواست وام بانکی تعلق دارند [1].

تنها اطلاعاتی که طبیعت فرآیند موجود می‌باشد، به شرح زیر است؛ در ابتدا مشتری وجه خاصی را انتخاب کرده و سپس درخواست خود را از طریق سایت بانک به صورت آنلاین ثبت می‌نماید. برخی از وظایف در این فرآیند به شکل خودکار انجام می‌شوند. به عنوان مثال، می‌توان بررسی نمود که آیا یک درخواست از سمت مشتری برای أخذ اعتبار بانکی، واجد شرایط است یا خیر. انواع مختلفی از وظایف در این فرآیند به کار گرفته شده‌اند. وظایفی که بدون دخالت یا همراه با دخالت نیروی انسانی صورت می‌گیرد.

شناسایی فرآیندهایی که منجر به ارائه‌ی یک محصول یا خدمت به مشتری می‌شوند، امر بسیار مهمی است و یک زمینه تحقیقاتی فعال در جامعه علمی فرض می‌گردد، به خصوص در حوزه مدیریت فرآیندهای کسب‌وکار [2].

۱-۱- مدیریت فرآیندهای کسب‌وکار

مدیریت فرآیندهای کسب‌وکار به عنوان مجموعه‌ای از تکنیک‌های بهینه‌سازی فرآیندهای کسب‌وکار معرفی شده است که تضمین می‌کند با تشخیص وظایف تکراری، چرخه‌ها یا مسیرهای کم‌تکرار و غیرسودآور، منجر به افزایش بهره‌وری، کارایی و کاهش هزینه‌های عملیاتی شرکت گردد. در این شرایط، یک فرآیند کسب‌وکار به عنوان مجموعه‌ای از وظایف تعریف می‌شود که به ترتیب و با یک توالی مشخص اجرا می‌شوند تا در نهایت به تولید محصول یا ارائه خدمت به مشتری منتهی شوند [3,4].

یکی از تکنیک‌های پرکاربرد در این حوزه، فرآیندکاوی است. در حقیقت فرآیندکاوی به عنوان یک ابزار مفید، امکان تجزیه و تحلیل خودکار فرآیندهای کسب‌وکار را بر اساس رکوردهای ثبت‌شده از رویدادها فراهم می‌کند.

به جای مدلسازی یک فرآیند بر اساس آنچه باید در حقیقت اتفاق بیفتد، فرآیندکاوی به جمع‌آوری اطلاعات رویدادهایی که در طول فرآیند گردش کار انجام می‌شوند، مشغول است و این داده‌ها را در فرمت‌های ساختاری به نام لاگ‌های رویداد ذخیره می‌کند تا بر اساس آن مدل فرآیندی را کشف نماید [5]. به بیانی دیگر فرآیندکاوی کاشف آن چیزی است که در حقیقت رخ می‌دهد. در هنگام جمع‌آوری این اطلاعات، فرض می‌شود که [6]:

- هر رویداد به یک وظیفه در فرآیند اشاره دارد؛
- هر رویداد یک نمونه از گردش کار را نشان می‌دهد؛
- از آن جایی که رویدادها بر اساس زمان اجرای خود ثبت می‌شوند، فرض می‌گردد که مرتب شده‌اند؛

بنابراین ترتیب فعالیت‌ها مشخص‌کننده نوع روابط علی و معلولی بین‌شان نیز خواهد بود که در مدل‌های کشف و استخراج شده از گزارش رویداد[1] قابل مشاهده است [7].

[1]. Event Log

۱-۲- داده‌های گم‌شده

رکوردهای ثبت شده در گزارش رویداد، منبع اصلی کشف فرآیندهای کسب‌وکار به حساب می‌آیند. با این حال، معمول است که این داده‌ها ناقص و یا همراه با مقدار زیادی اطلاعات گم‌شده باشند، به عنوان مثال فراموشی منابع انسانی در خصوص ثبت وظایف خود یا خرابی سیستم و سایر موارد. معمولاً روش‌های آماری برای رفع مشکل داده‌های گم‌شده به کار گرفته می‌شوند. با این حال، بیشتر روش‌های آماری نیاز به یک مجموعه داده کامل یا حداقل یک مجموعه داده کافی و قوی دارند تا پیش‌بینی‌های دقیقی انجام دهند [8]. عدم وجود داده‌های کامل منجر به کاهش شدید دقت می‌شود و نتایج مدل‌های آماری را به مخاطره می‌اندازد.

۲- مطالعه‌ی موردی

گزارش رویداد مورد استفاده در این پروژه متعلق به بانکی در هلند است و فرآیند درخواست وام بانکی را نشان می‌دهد. این مجموعه‌ی داده در سال ۲۰۱۲ ارائه شده است. توضیحات فرآیند به شرح ذیل می‌باشد [1].

- درخواست وام بانکی با ورود مشتری به صفحه وب بانک مذکور آغاز می‌شود و مشتری مقدار مشخصی پول را تحت عنوان وام دریافتنی انتخاب و سپس درخواست خود را ارسال می‌کند.
- در گام بعدی، برخی از وظایف به شکل خودکار انجام می‌شوند. بدین معنا که بررسی می‌گردد که آیا درخواست ثبت شده واجد شرایط است یا خیر.
- اگر واجد شرایط باشد، پیشنهادی از طریق ایمیل (یا تلفن) به مشتری ارسال می‌شود. ارسال پیشنهاد در فرآیند درخواست وام بانکی به معنای ارائه‌ی یک پیشنهاد وام به مشتری است. این پیشنهاد معمولاً شامل جزئیاتی است نظیر مبلغ وام، نرخ بهره، مدت زمان بازپرداخت و سایر شرایط مربوطه. این پیشنهاد براساس اطلاعاتی که مشتری در فرم درخواست ارائه کرده است، تهیه می‌گردد.
- پس از ارسال پیشنهاد وام بانکی توسط بانک، اگر مشتری آن را پذیرفت، فرآیند ادامه می‌یابد. ولاغیر دلیلی بر ادامه فرآیند ثبت درخواست وام بانکی نخواهد بود.
- پس از دریافت تأیید مشتری، ارزیابی‌های لازم صورت خواهد گرفت. در این مرحله بانک تصمیم می‌گیرد که تا چه اندازه متقاضی وام بانکی می‌تواند به تعهدات خود عمل نمایند. این تصمیم بر اساس تحلیل آماری از سوابق مالی و اعتباری متقاضی اتخاذ می‌گردد. به عبارت دیگر، ارزیابی پیشنهاد به معنای بررسی توانایی مشتری در بازپرداخت وام است.
- در صورت کامل نبودن اطلاعات سوابق مشتری برای پیش بردن تصمیم‌گیری مذکور، پیشنهاد به مشتری باز می‌گردد تا با جمع آوری کلیه اطلاعات مورد نیاز مجدداً ارزیابی صورت گیرد.
- در آخر ارزیابی نهایی انجام و سپس درخواست تایید می‌شود.

همچنین لازم به ذکر است که فرآیند از سه گروه مختلف رویداد تشکیل شده است. حرف اول هر وظیفه مربوط به یک شناسه از زیرفرآیندی است که به آن تعلق دارد [1]:

- وظایفی که با حرف ***A*** شروع می‌شوند، به وضعیت‌هایی از درخواست وام بانکی اشاره دارد که به صورت خودکار انجام می‌شود.
- وظایفی که با حرف ***O*** شروع می‌شوند به پیشنهاداتی بر می‌گردد که به مشتری ابلاغ می‌شود. از مجموعه‌ی داده مشخص نیست که آیا این وظایف به صورت خودکار توسط یک برنامه صورت می‌گیرد یا این که با دخالت نیروی انسانی همراه است.
- وظایفی که با حرف ***W*** شروع می‌شوند بر حالت‌هایی دلالت دارد که در خلال فرآیند رخ می‌دهند. این رویدادها مبتنی بر دخالت نیروی انسانی هستند.

۲-۱- مجموعه‌ی رویدادها

گزارش رویداد یک مجموعه‌ی داده‌ی ساختاریافته است که به مقدار قابل‌توجهی پردازش برای شناسایی و استخراج اطلاعات در راستای تجزیه و تحلیل فرآیند احتیاج دارد. همچنین خلاصه‌ی تمامی وظایف موجود به ترتیب در جداول ۱ تا ۳ فراهم شده است.

جدول ۱ - وضعیت درخواست وام بانکی (به صورت خودکار)

رویداد	تعداد دفعات رخداد	توضیحات
A-SUBMITTED	۱۳۰۸۷	رویدادهای آغازین. کلیه درخواست‌های ثبت‌شده در مجموعه‌ی داده با این گروه از رویدادها آغاز شده‌اند. این وظایف مربوط به اقدام مشتری برای ارسال درخواست أخذ وام بانکی است.
A-PARTLYSUBMITTED	۱۳۰۸۷	
A-PREACCEPTED	۷۳۶۷	این رویداد نشان می‌دهد که درخواست‌هایی است که به ثبت رسیده ولیکن هنوز پذیرفته نشده است، چرا که به اطلاعات بیشتری نیاز دارد.
A-ACCEPTED	۵۱۱۳	درخواست پذیرفته شده و آماده رفتن به مرحله نهایی است. با این حال، هنوز هم ممکن است به برخی اطلاعات اضافی از مشتری نیاز داشته باشد.
A-FNIALIZED	۵۰۱۵	درخواست ارسال شده به طور کامل پذیرفته شده و آماده ارزیابی است.
A-CANCELLED	۲۸۰۷	رویدادهای پایانی برای درخواست‌هایی که ناموفق بوده‌اند. هرچند چندان واضح نیست فرق بین این دو رویداد در چیست.
A-DECLINED	۷۶۳۵	
A-APPROVED	۲۲۴۶	رویدادهای پایانی برای درخواست‌هایی که موفقیت‌آمیز بوده‌اند.
A-REGISTERED	۲۲۴۶	
A-ACTIVATED	۲۲۴۶	

جدول ۲ - پیشنهادات ابلاغ شده به مشتری (به صورت خودکار یا توسط نیروی انسانی)

رویداد	تعداد دفعات رخداد	توضیحات
O-CREATED	۷۰۳۰	
O-SELECTED	۷۰۳۰	پیشنهاد مناسب برای متقاضی ایجاد شده و برای مشتری منتخب ارسال می‌گردد.
O-SENT	۷۰۳۰	
O-SENTBACK	۳۴۵۴	پاسخ مشتری به پشنهاد دریافت شده است.
O-ACCEPTED	۲۲۴۳	رویداد پایانی برای پیشنهادهایی که موفقیت‌آمیز بوده و همچنین به توافق طرفین رسیده است.
O-CANCELLED	۳۶۵۵	رویدادهای پایانی برای پیشنهادهایی که ناموفق بوده‌اند. لغو صورت گرفته می‌تواند از سمت
O-DECLINED	۸۰۲	مشتری یا موسسه‌ی بانکی بنابر هر دلیلی رخ داده باشد.

جدول ۳ - وظایف نیروی انسانی در طول اجرای فرآیند

رویداد	تعداد دفعات رخداد	توضیحات
W-CALLING AFTER SENT OFFERS	۵۲۰۱۶	هر زمان که پیشنهادی برای مشتری ارسال می‌گردد، فعال می‌شود.
W-ASSESING THE APPLICATION	۲۰۸۰۹	شرایط درخواست ارسال شده از سمت مشتری مورد ارزیابی قرار می‌گیرد.
W-FILLING IN INFORMATION	۵۴۸۵۰	پیگیری اطلاعات برای درخواست‌های (PREACCEPTED).
W-FIXING INCOMING LEAD	۱۶۵۶۶	در ابتدای فرآیند درخواست، اگر مشتری اطلاعات مورد نیاز را پر نکرده باشد به وقوع می‌پیوندد.
W-CALLING TO ADD MISSING INFORMATION	۲۵۱۹۰	بعد از ارزیابی درخواست مشتری نیاز به اطلاعات اضافی وجود دارد.
W-RATE FRAUD	۶۶۴	بعد از ارزیابی درخواست، موارد مشکوک به کلاهبرداری بررسی می‌شوند.
W-CHANGE CONTRACT DETAILS	۰	زمانی که نیاز به تغییرات مفاد قرارداد باشد، این رویداد رخ می‌دهد.

لازم به ذکر است، گزارش رویداد فرآیند، اطلاعات مربوط به توالی زمانی فعالیت‌ها را نیز در بر می‌گیرد. سه شناسه‌ی زیر می‌توانند بینش دقیق‌تری از مدت زمان انجام فعالیت‌ها ارائه دهند، بدین صورت که؛

- آغاز شده[1]: شروع کار نیروی انسانی را نشان می‌دهد. (W)
- خاتمه یافته[2]: بیان‌گر به پایان رسیدن یک فعالیت است. (A - O)
- برنامه‌ریزی شده[3]: کارهایی که انجام آن‌ها به یک زمان (تاریخ) مشخصی موکول می‌گردد، بدین صورت علامت‌گذاری می‌شوند.

[1]. Start
[2]. Complete
[3]. Schedule

فصل دوم: تحلیل‌های توصیفی

۱- توصیف مجموعه‌ی رویدادها

در ابتدای امر لازم است توضیحاتی در خصوص گزارش رویداد فرآیند درخواست وام بانکی ارائه گردد. از آن‌جایی گزارش رویدادها بر مبنای فرمت *pm4py.objects.log.obj.EventLog* خوانایی و درک چندان زیادی ندارد، بهتر است در قالب یک *Data Frame* نمایش داده شود. جدول ۴، چند سطر ابتدایی گزارش رویداد فرآیند را نشان می‌دهد.

جدول ۴ - گزارش رویداد اولیه فرآیند درخواست وام بانکی

org:resource	lifecycle:transition	concept:name	time:timestamp	case:REG_DATE	case:concept:name	case:AMOUNT_REQ
112	COMPLETE	A_SUBMITTED	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:38:44.546000+00:00	173688	20000
112	COMPLETE	A_PARTLYSUBMITTED	2011-10-01 00:38:44.880000+00:00	2011-10-01 00:38:44.546000+00:00	173688	20000
112	COMPLETE	A_PREACCEPTED	2011-10-01 00:39:37.906000+00:00	2011-10-01 00:38:44.546000+00:00	173688	20000
112	SCHEDULE	W_Completeren aanvraag	2011-10-01 00:39:38.875000+00:00	2011-10-01 00:38:44.546000+00:00	173688	20000
NaN	START	W_Completeren aanvraag	2011-10-01 11:36:46.437000+00:00	2011-10-01 00:38:44.546000+00:00	173688	20000
…	…	…	…	…	…	…

در ادامه برای بهبود ساختار گزارش رویداد مربوطه، تغییراتی در اسم و ترتیب ستون‌ها اعمال می‌گردد که در جدول ۵، قابل مشاهده می‌باشد.

جدول ۵ - گزارش رویداد ثانویه فرآیند درخواست وام بانکی

Case_ID	Activity	Transition	Resource	Start_Timestamp	Complete_Timestamp	Amount_Request
173688	A_SUBMITTED	COMPLETE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:38:44.546000+00:00	20000
173688	A_PARTLYSUBMITTED	COMPLETE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:38:44.880000+00:00	20000
173688	A_PREACCEPTED	COMPLETE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:39:37.906000+00:00	20000
173688	W_Completeren aanvraag	SCHEDULE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:39:38.875000+00:00	20000
173688	W_Completeren aanvraag	START	NaN	2011-10-01 00:38:44.546000+00:00	2011-10-01 11:36:46.437000+00:00	20000
…	…	…	…	…	…	…

ستون‌های گزارش رویداد فرآیند بیانگر موارد زیر هستند:

- *Case_ID*: نشان‌دهنده‌ی شماره درخواست مشتری است که کدی منحصر به فرد تلقی می‌گردد. لازم به ذکر است در این گزارش، کلیه‌ی مقاطع زمانی مربوط به یک درخواست، به ترتیب و پشت سر هم ثبت شده و سپس رویدادهای مربوط به درخواست بعدی درج می‌شود. بنابراین چند سطر ابتدایی مشاهده شده در جدول ۵، همگی متعلق به درخواست یک مشتری به شماره‌ی *<173688>* می‌باشد.
- *Activity*: کلیه فعالیت‌های صورت گرفته در حین فرآیند را برای یک درخواست نمایش می‌دهد.
- *Transition*: اطلاعات مربوط به توالی زمانی فعالیت‌ها را نیز در بر می‌گیرد که در قالب سه شناسه‌ی *Start*، *Schedule* و *Complete* بینش دقیق‌تری از مدت زمان انجام فعالیت‌ها ارائه دهند که پیش‌تر جزئیات آن ارائه گردیده است.
- *Resource*: هر یک از فعالیت‌های لازمه در خلال فرآیند توسط منبع خاصی انجام می‌شود که به صورت یک کد قابل پیگیری است.
- *Start_Timestamp*: مقادیر مربوط به این ستون برای کلیه‌ی رویدادهای یک درخواست یکسان بوده و به صورت کلی بیانگر مقطع زمانی و تاریخی است که درخواست مشتری به صورت آنلاین به ثبت رسیده است.
- *Complete_Timestamp*: زمان به پایان رسیدن و خاتمه‌ی هر رویداد از طریق مقادیر ذکر شده در این ستون قابل مشاهده می‌باشد.
- *Amount_Request*: مقدار وام درخواستی را برای هر مشتری مشتری نشان می‌دهد، بنابراین توقع می‌رود برای کلیه لاگ‌های ثبت شده مرتبط با یک درخواست منحصر به فرد، ستون میزان وام درخواستی مقادیر یکسانی به خود بگیرد.

۱-۱- تحلیل ساختاری

در این بخش تلاش می‌گردد با ارائه‌ی جداول و نمودارهایی، تسهیل فهم اطلاعات درج شده در گزارش رویداد میسر گردد.

اولین نکته‌ی حائز اهمیت در مواجه با گزارش رویداد یک فرآیند، پیدا نمودن تعداد کل درخواست‌های ثبت شده می‌باشد. جدول ۶، شرحی از درخواست‌ها و تعداد رکوردهای ثبت شده در هر درخواست را نشان می‌دهد.

جدول ۶ - تعداد درخواست‌ها و رکوردهای ثبت شده هر درخواست

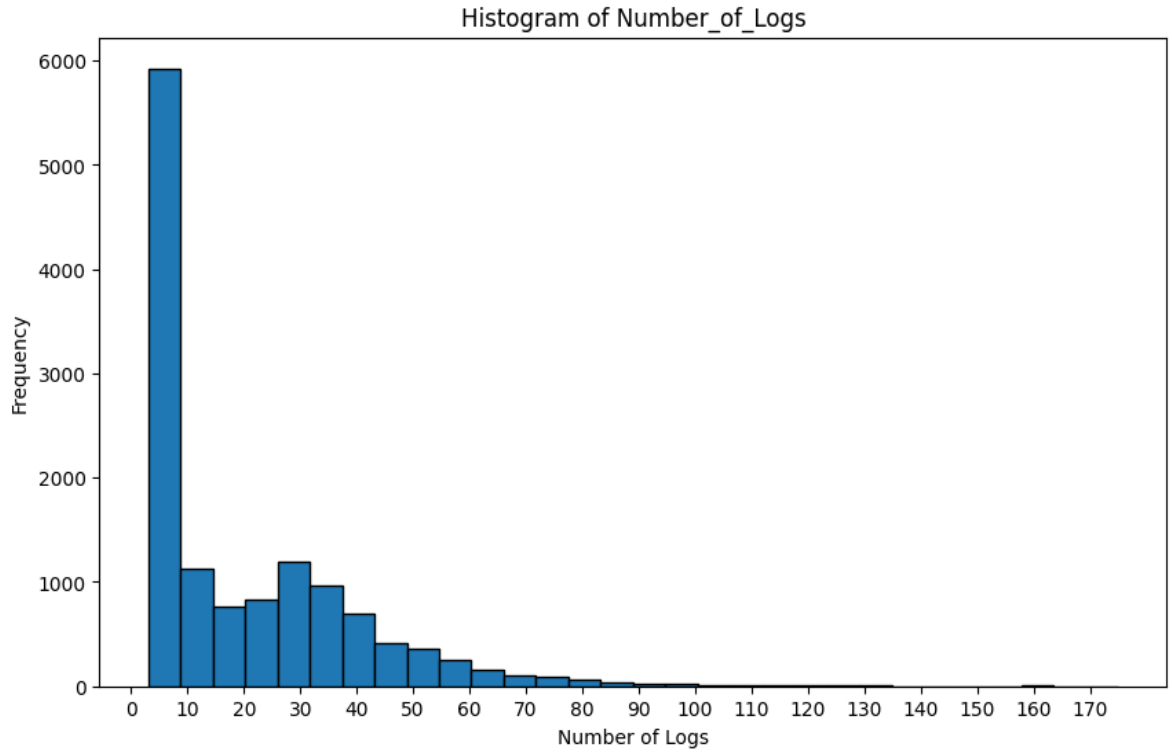
No.	Case_ID	Number of Logs
0	194055	3
1	213255	3
2	180989	3
3	181007	3
4	210647	3
…	…	…
13082	181799	161
13083	198232	163
13084	183175	167
13085	195247	170
13086	185548	175
Minimum Case_ID: 173688 **Maximum Case_ID:** 214376		

همان‌طور که قابل مشاهده است، در مجموع ۱۳۰۸۷ درخواست در خلال این فرآیند به ثبت رسیده است که با توجه به خروجی گرفته شده، کمترین تعداد لاگ ثبت شده برای هر درخواست ۳ و بیشترین تعداد آن ۱۷۵ می‌باشد.

به عبارت دیگر، تعداد لاگ ثبت شده نشان از همه‌ی فعالیت‌هایی است که از ابتدا تا انتهای رسیدگی به یک درخواست انجام شده‌اند. همچنین شکل ۱، هیستوگرام تعداد رکوردهای ثبت‌شده در هر درخواست را نشان می‌دهد.

مطابق با کمترین و بیشترین شناسه‌ی درخواست مشتری به شرح جدول ۶، توقع می‌رود ۴۰۶۸۸ در خواست به ثبت رسیده باشد، ولیکن تعداد کل ۱۳۰۸۷ درخواست ثبت‌شده، دو فرضیه زیر را نشان می‌دهد:

- گزارش رویداد فعلی، تنها یک سوم جزئیات درخواست‌های به ثبت رسیده را شامل می‌شود.
- یا این‌که به صورت کلی شناسه‌های درخواست به صورت مضرب سوم (تقریبی) به هر مشتری تخصیص داده می‌شوند.

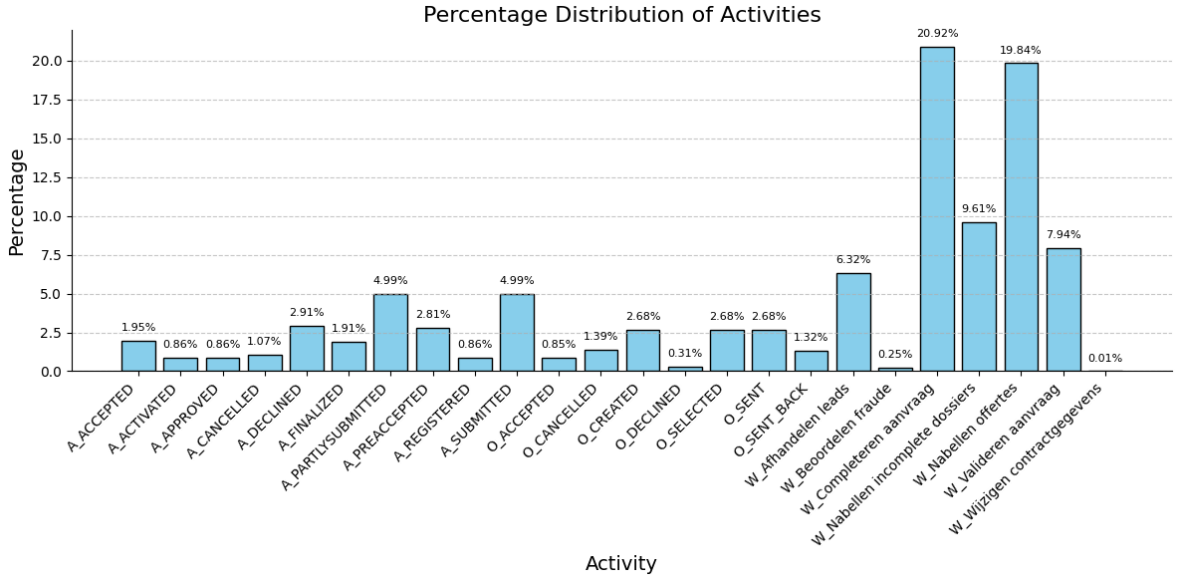


شکل ۱ - هیستوگرام تعداد رکوردهای ثبت شده هر درخواست

از طرفی بهتر است بررسی شود، رکوردهای ثبت شده در گزارش رویداد شامل چه فعالیت‌هایی می‌گردد. در این راستا جدول ۷، کلیه‌ی فعالیت‌های به کار رفته در فرآیند و درصد رخداد آن‌ها را نشان می‌دهد. همچنین شکل ۲، هیستوگرام درصد توزیع فعالیت‌های را نشان می‌دهد.

جدول ۷ – فعالیت‌های فرآیند درخواست وام بانکی

No.	Unique_Values	Percentage
0	A_ACCEPTED	1.950
1	A_ACTIVATED	0.857
2	A_APPROVED	0.857
3	A_CANCELLED	1.071
4	A_DECLINED	2.912
5	A_FINALIZED	1.913
6	A_PARTLYSUBMITTED	4.991
7	A_PREACCEPTED	2.810
8	A_REGISTERED	0.857
9	A_SUBMITTED	4.991
10	O_ACCEPTED	0.855
11	O_CANCELLED	1.394
12	O_CREATED	2.681
13	O_DECLINED	0.306
14	O_SELECTED	2.681
15	O_SENT	2.681
16	O_SENT_BACK	1.317
17	W_Afhandelen leads	6.318
18	W_Beoordelen fraude	0.253
19	W_Completeren aanvraag	20.919
20	W_Nabellen incomplete dossiers	9.607
21	W_Nabellen offertes	19.838
22	W_Valideren aanvraag	7.936
23	W_Wijzigen contractgegevens	0.005



شکل ۲ – هیستوگرام درصد توزیع فعالیت‌های گزارش رویداد

کلیه مسیرهای منحصر به فرد موجود در فرآیند با فعالیت *A_SUBMITTED* آغاز شده است. از طرفی جدول ۸ برای مشخص نمودن رویداهای پایانی ممکن در مسیرهای فرآیند ارائه شده است.

جدول ۸ - رویدادهای پایانی فرآیند درخواست وام بانکی

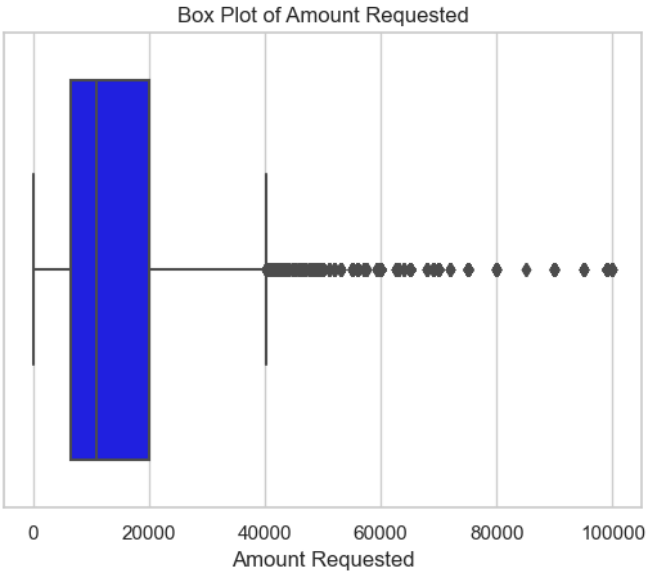
End_Event	Min_Amount_Request	Mean_Amount_Request	Max_Amount_Request	Ratio
W_Valideren aanvraag	25	15889.02	99000	20.99
W_Completeren aanvraag	10	15802.73	99999	14.82
A_CANCELLED	0	14962.89	75000	5.00
W_Afhandelen leads	1	10864.57	99000	17.07
O_CANCELLED	2000	15943.6	90000	2.13
A_DECLINED	12	10868.21	99999	26.2
W_Beoordelen fraude	500	11012.7	49350	0.44
A_REGISTERED	20000	20000	20000	0.01
W_Nabellen incomplete dossiers	1000	16811.28	65000	3.45
W_Nabellen offertes	500	14933.3	99000	9.86
W_Wijzigen contractgegevens	6000	12000	25000	0.03

همچنین لازم است رکوردهای ثبت شده در گزارش رویداد را از منظر انواع مختلف *Transition*های ثبت شده بررسی نمود. بدین جهت جدول ۹، کلیه‌ی *Transition*های به کار رفته در فرآیند و درصد رخداد آن‌ها را نشان می‌دهد.

جدول ۹ - *Transition*های فرآیند درخواست وام بانکی

No.	Unique_Values	Percentage
0	COMPLETE	62.741
1	SCHEDULE	10.037
2	START	27.222

در ادامه منابع موجود در فرآیند بررسی می‌شوند. در کل ۶۸ منبع در فرآیند موجود است که کمترین و بیشترین کد تخصیص داده شده به این منابع به ترتیب ۱۱۲ و ۱۱۳۳۹ می‌باشد. همچنین درخصوص مقدار وام درخواستی نیز، کمترین و بیشترین میزان ثبت شده در درخواست‌ها به ترتیب صفر و ۹۹۹۹۹ می‌باشد. برای ارائه‌ی درک بهتری از مقدار وام درخواستی مشتریان، نمودار جعبه‌ای مطابق شکل ۳ ارائه شده است و اطلاعات لازم جهت توصیف نمودار مذکور به شرح جدول ۱۰ می‌باشد.



شکل ۳ - نمودار جعبه‌ای مقدار وام درخواستی مشتریان

جدول ۱۰ - اطلاعات آماری نمودار جعبه‌ای مقدار وام درخواستی مشتریان

Parameter	Value
Mean	15586.795381
Standard Deviation	12381.430915
Minimum	0.000000
First Quartile	6500.000000
Second Quartile	11000.000000
Third Quartile	20000.000000
Maximum	99999.000000

۱-۲- پر کردن مقادیر گم‌شده

داده کاوی[1] به عنوان یک وظیفه مهم و چالش‌برانگیز در بسیاری از مسائل زندگی روزمره شناخته شده است. برای تحلیل داده‌های بزرگ، مجموعه‌ی داده برای یک مسئله با یک هدف مشخص جمع‌آوری می‌شود. با این حال، در عمل، مجموعه داده‌ی جمع‌آوری‌شده معمولاً شامل نسبتی از داده‌های ناقص است که یک یا چند مقدار ویژگی دارای مقادیر گم‌شده هستند. بسیاری از دلایل ناقص بودن مجموعه‌ی داده از منابع مختلفی ناشی می‌شود، از جمله سیستم پایگاه داده به طور مستقیم، شبکه، ورودی‌های داده‌ی نادرست و غیره [9].

وقتی که مجموعه‌ی داده حاوی مقادیر بسیار کمی از داده‌های گم‌شده است، به عنوان مثال نرخ گم‌شدگی کمتر از ۱۰٪ یا ۱۵٪ برای کل مجموعه‌ی داده است، می‌توان به سادگی داده‌های گم‌شده را از مجموعه داده حذف کرد بدون این‌که تأثیر قابل‌توجهی بر نتیجه‌ی نهایی کاوش یا تحلیل داشته باشد. با این حال، وقتی نرخ گم‌شدگی بیشتر از ۱۵٪ است، نظارت دقیق در مورد نحوه‌ی رفتار با داده‌های گم‌شده ضروری است. باید توجه داشت که این بدان معنا نیست که هر مجموعه داده مسئله در حوزه این نوع قانون را دنبال می‌کند. بدین منظور روش پر کردن مقادیر گم‌شده[2]، ارائه می‌گردد. رویکرد مذکور روشی است که بیشترین کاربرد را در برخورد با مشکل داده‌های ناقص دارد. به طور کلی، فرآیندی است که از برخی تکنیک‌های آماری یا یادگیری ماشین[3] برای جایگزینی داده‌های گم‌شده با مقادیر جایگزین استفاده می‌کند.

---
[1]. Data Mining
[2]. Missing Value Imputation
[3]. Machine Learning Methods

۱-۲-۱- دسته‌بندی انواع داده‌ی گم‌شده

در این بخش یک دسته‌بندی کلی برای انواع داده‌های گم‌شده ارائه می‌گردد که به شرح ذیل است؛

- **داده‌ی گم‌شده‌ی کاملاً تصادفی**[1]

  در این دسته از داده‌های تصادفی، مقادیر گم‌شده به صورت کاملاً تصادفی در سراسر مجموعه‌ی داده پخش می‌شوند و ارتباطی بین این مقادیر با سایر متغیرهای مجوعه‌ی داده یافت نمی‌شود. به عنوان نمونه، می‌توان به مقدار از دست رفته در یک پرسش‌نامه اشاره کرد که صرفاً به علت عدم توجه از مقداردهی به آن صرف نظر شده است.

- **داده‌ی گم‌شده‌ی تصادفی**[2]

  در این دست از مقادیر گم‌شده، احتمال خالی بودن یک مقدار به وابستگی بین متغیرهای مشاهده‌شده ارتباط دارد اما به مقادیر گم‌شده‌ی خود وابسته نیست. به عنوان مثال، اگر در پرسش‌نامه‌ای پاسخ‌دهنده به یک سوال حساس و یا شخصی پاسخ ندهد به احتمال بالا به سایر پرسش‌های مرتبط به این سوال نیز پاسخ نمی‌دهد.

- **داده‌ی گم‌شده‌ی غیرتصادفی**[3]

  در این دست از مقادیر گم‌شده، احتمال گم‌شدن یک مقدار وابستگی بالایی به عدم مشاهده‌ی برخی دیگر از متغیرها و همچنین مقادیر خود رکورد دارد. به عنوان نمونه، می‌توان به فردی در پرسش‌نامه اشاره کرد که به علل شخصی مثل عدم اعتماد تمایلی به پاسخ‌دهی پرسش‌ها ندارد.

---

[1]. Missing Completely at Random (MCAR)
[2] Missing at Random (MAR)
[3] Missing Not at Random (MNAR)

۱-۲-۲- کارآمدترین روش‌های پر کردن مقادیر گم‌شده

پراستفاده‌ترین روش‌های پر کردن داده‌های گم‌شده به صورت زیر قابل تفکیک و گروه‌بندی است.

- **پر کردن مقادیر گم‌شده با میانگین، میانه یا مد:**

این روش که یکی از ساده‌ترین روش‌های پر کردن مقادیر گم‌شده یا شاید ساده‌ترین روش موجود است، با استفاده از فرضیه‌های ساده‌ی آماری به پر کردن مقادیر گم‌شده می‌پردازد و به طور کلی مناسب دسته‌ی داده‌های گم‌شده‌ی کاملاً تصادفی است که به وسیله‌ی میانگین، میانه و یا مد به پر کردن داده‌های گم‌شده می‌پردازد. در این رویکرد در صورتی که داده‌ها عددی باشند، با استفاده از میانگین (برای توزیع‌های متقارن) یا میانه (برای توزیع‌های چوله) به پر کردن داده‌ها پرداخته می‌شود. همچنین در صورتی که متغیر مورد نظر از جنس کیفی باشد، داده‌ها به وسیله‌ی مد مورد ارزیابی و اصلاح قرار می‌گیرند. در این روش همچنین می‌توان از رویکردهای خلاقانه استفاده کرد تا بتوان به‌وسیله‌ی دسته‌بندی حالت‌ها با دقت بالاتری به پرکردن مقادیر پرداخت.

- **پر کردن مقادیر گم‌شده با استفاده از تکنیک‌های شبیه‌سازی:**

این راهکار که از رویکرد فوق پیچیدگی بیشتری دارد، مناسب دسته‌ی داده‌های گم‌شده‌ی تصادفی است. با توجه به ارتباط حداقلی بین متغیر دارای مقادیر گم‌شده و سایر متغیرها می‌توان بر أساس رابطه‌ی احتمالی بین این مقادیر قواعد و روابطی تعریف کرد تا بتوان بر أساس این أصول به صورت تصادفی این مقادیر خالی را پر کرد.

- **پر کردن مقادیر گم‌شده با استفاده از تکنیک‌های خوشه‌بندی:**

در این روش که پایه‌ی آن مبتنی بر خوشه‌بندی است، تلاش می‌شود تا بر اساس روابط تعریف شده در الگوریتم‌های خوشه‌بندی و به وسیله‌ی معیارهای مختلف نزدیکی و شباهت که در آن‌ها تعریف می‌شود به پر کردن مقادیر گم‌شده پرداخت. این رویکرد که خود شامل روش‌های مختلفی است، مبتنی بر شاخص‌های مختلف و همچنین ساختمان داده‌ی متغیرها بر اساس قرارگیری رکورد در یک خوشه به پر کردن مقادیر آن می‌پردازد. با توجه به تفسیرات این رویکرد می‌توان آن را برای دسته داده‌های گم‌شده‌ی غیرتصادفی استفاده کرد.

- **پرکردن مقادیر گم‌شده با استفاده از تکنیک‌های دسته‌بندی و رگرسیون:**

این رویکرد که پس‌زمینه‌ی آن از جنس روش‌های حل یادگیری ماشین است، با توجه به ساختار داده به پر کردن مقادیر گم‌شده با استفاده از روش‌های رگرسیون و دسته‌بندی می‌پردازد. الگوریتم حل این دسته از روش‌ها بر اساس مقداردهی متغیر دارای مقدار گم‌شده بر اساس سایر متغیر هاست که با این اوصاف از جنس روش‌های پیش‌بینی است. با توجه به این که پایه و اساس این روش بر حسب رابطه‌ی بین متغیرهاست، می‌توان از این روش برای پر کردن مقادیر گم‌شده‌ی دسته‌ی داده‌های تصادفی استفاده کرد.

۱-۲-۳- شرح رویکرد اعمال‌شده بر گزارش رویداد

حال که به اختصار شرحی از انواع مقادیر گم‌شده و روش‌های پر کردن آن ارائه شد، می‌توان به گزارش آن‌چه در این مجموعه‌ی داده انجام شده است، پرداخت. با توجه به انواع داده‌ی گم‌شده که پیشتر توضیح داده شد، در گام اول به استخراج درصد مقادیر غیرگم‌شده برای هر ویژگی در مجموعه‌ی داده پرداخته که در ادامه می‌توان نتایج حاصل را مطابق جدول ۱۱، مشاهده نمود:

جدول ۱۱ - پیدایش مقادیر از دست رفته‌ی گزارش رویداد

Feature	Non-Missing Percentage
Case_ID	100.00
Activity	100.00
Transition	100.00
Resource	100.00
Strat_Timestamp	100.00
Complete_Timestamp	100.00
Amount_Request	100.00
Resource	93.13

نتایج حاکی از آن است که تنها منابع به کار گرفته شده در فرآیند، دارای مقادیر از دست رفته است. لازم به ذکر است که این بدان معنا نیست که سایر ویژگی‌ها دارای مقادیر گم‌شده نیستند، بلکه ممکن است در ستون‌های دیگر مجموعه‌ی داده نیز شاهد مقادیر گم‌شده باشیم. همانند حالتی که مقادیر به صورت خط تیره، ۱- یا سایر مقادیر مشابه در مجموعه‌ی داده موجود باشد، که در گام اول این پروژه این امر مورد بررسی قرار گرفته است. با توجه به شرح فوق، در وهله‌ی اول به پر کردن مقادیر گم‌شده‌ی منابع پرداخته شد که بر اساس بیشترین تکرار هر منبع (مد) در هر فعالیت از لاگ است. همچنین مورد بعد که قابل توجه است، مشاهده‌ی مقادیر اعشار از دست رفته‌ی زمان‌های شروع و پایان (میلی ثانیه) در برخی رکوردها است که مجموعاً ۷۷۰ لاگ را تشکیل می‌دهند. این چالش تنها مربوط به یک فعالیت خاص نبوده و در بسیاری از فعالیت‌های گزارش رویداد چنین اتفاقی رخ داده است. بنابراین فعالیت خاصی دچار مشکل ثبت رکورد با چنین خطایی نمی‌باشد. از آن جا که برای تحلیل‌های آتی نیازی به این مقادیر اعشار نیست و همچنین پرکردن یا حذف آن‌ها احتمال خطا را افزایش می‌دهد، تصمیم بر حفظ این مقادیر شد.

۱-۳- درخت تصمیم‌گیری فرآیند

پیش از آن‌که درخت تصمیم‌گیری فرآیند در این بخش مورد بررسی قرار گیرد، لازم است تغییراتی بر روی فعالیت‌های فرآیند اعمال گردد تا سرعت پیگیری و تحلیل‌های مورد نیاز افزایش یابد. بدین جهت جدول ۱۲، برای تبدیل گزارش رویداد به گزارش جریان کار[1] ارائه شده است. کلیه فعالیت‌های صورت گرفته در فرآیند درخواست وام بانکی، به ترتیب حروف الفبا، نام‌گذاری شده‌اند.

جدول ۱۲ - گزارش جریان کار

No.	Activity	Letter
0	A_SUBMITTED	a
1	A_PARTLYSUBMITTED	b
2	A_PREACCEPTED	c
3	A_ACCEPTED	d
4	A_FINALIZED	e
5	A_APPROVED	f
6	A_REGISTERED	g
7	A_ACTIVATED	h
8	A_CANCELLED	i
9	A_DECLINED	j
10	O_SELECTED	k
11	O_CREATED	l
12	O_SENT	m
13	O_SENT_BACK	n
14	O_ACCEPTED	o
15	O_CANCELLED	p
16	O_DECLINED	q
17	W_Afhandelen leads	r
18	W_Completeren aanvraag	s
19	W_Nabellen offertes	t
20	W_Valideren aanvraag	u
21	W_Nabellen incomplete dossiers	v
22	W_Beoordelen fraude	w
23	W_Wijzigen contractgegevens	x

[1]. Workflow Log

با توجه به تغییرات اعمال شده لازم است چند سطر ابتدایی گزارش رویداد در قالب یک *Data Frame* جدید مطابق جدول ۱۳، مجدداً نمایش داده شود. همچنین جدول ۱۴ نیز به منظور تشریح کلیه واریانت‌های موجود در فرآیند و نمایش رویدادهای آغازین و پانانی آن‌ها ارائه شده است.

جدول ۱۳ – گزارش رویداد بر اساس گزارش جریان کار

Case_ID	Activity	Transition	Resource	Start_Timestamp	Complete_Timestamp	Amount_Request
173688	a	COMPLETE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:38:44.546000+00:00	20000
173688	b	COMPLETE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:38:44.880000+00:00	20000
173688	c	COMPLETE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:39:37.906000+00:00	20000
173688	s	SCHEDULE	112	2011-10-01 00:38:44.546000+00:00	2011-10-01 00:39:38.875000+00:00	20000
173688	s	START	NaN	2011-10-01 00:38:44.546000+00:00	2011-10-01 11:36:46.437000+00:00	20000
…	…	…	…	…	…	…

جدول ۱۴ – مسیرهای منحصر به فرد فرآیند

Case_ID	Variant	Start Event	End Event
173688	abcssdkelmtsttttnutugfohu	a	u
173691	abcsssdeklmtstttkplmtttttnutuuuuuofghu	a	u
173694	abcsssssssdeklmtstkplmttttttttkplmttttttttt...	a	x
173697	abj	a	j
173700	abj	a	j
…	…	…	…

اکنون لازم است توضیحاتی در خصوص نحوه ترسیم درخت تصمیم‌گیری فرآیند داده شود. رویکردی که بدین منظور در پیش گرفته شده است، به شرح زیر است؛

در ابتدای امر کافی است فعالیت‌هایی را که در حالت‌های تصمیم‌گیری درخواست وام بانکی یا اضافه برداشت موثر هستند را غربال نمود تا ادامه‌ی رویکرد قابل تشریح باشد. در این راستا نکات ذیل قابل توجه هستند:

- دو فعالیت *A_SUBMITTED* و *A_PARTLYSUBMITTED* در کلیه مسیرهای فرآیند به ترتیب و پشت هم ظاهر شده‌اند. بدین جهت دو فعالیت مذکور با یکدیگر ادغام شده و با حرف *a* نمایش داده می‌شوند، چرا که فعالیت دوم همیشه بلافاصله بعد از فعالیت نخست روی می‌دهد.
- سه فعالیت *A_REGISTERED*، *A_APPROVED* و *A_ACTIVATED* نیز در کلیه مسیرهای فرآیند به تریتب و پشت هم ظاهر شده‌اند و هیچ تفاوتی بین آن‌ها نمی‌باشد. بنابراین فعالیت‌های فوق با یکدیگر ادغام شده و تحت عنوان *g* قابل نمایش هستند.
- در آخر نیز سه فعالیت *O_CREATED*، *O_SELECTED* و *O_SENT* به ترتیب و پشت سر هم در گزارش رویداد ظاهر شده، بنابراین مشابه آن‌چه در خصوص دو گروه قبلی ذکر شد، برای فعالیت‌های این دسته هم ادغامی به نام *l* صورت می‌گیرد.

جدول ۱۵ فعالیت‌های فیلترشده‌ی تاثیرگذار بر وضعیت‌های درخواست وام بانکی را به صورت خلاصه نشان می‌دهد. در گام بعدی، مطابق با جدول ۱۶، تنها فعالیت‌های تأثیرگذار هر مسیر حفظ و سایر فعالیت‌ها حذف می‌گردند.

در حقیقت به تعبیری می‌توان فرض کرد رسم یک درخت تصمیم‌گیری برای حالت‌های مختلف درخواست وام بانکی و بررسی تعداد دفعات رخداد هر حالت، مشابه طی نمودن الگوریتم کشف فرآیند بر اساس گزارش رویدادی است که تنها از مسیرهایی شامل فعالیت‌های تأثیرگذار بر حالت‌های تصمیم‌گیری تشکیل شده است.

جدول ۱۵ - فعالیت‌های تأثیرگذار بر حالت‌های تصمیم‌گیری

No.	Activity	Letter
0	A_SUBMITTED A_PARTLYSUBMITTED	a
1	A_PREACCEPTED	c
2	A_ACCEPTED	d
3	A_FINALIZED	e
4	A_APPROVED A_REGISTERED A_ACTIVATED	g
5	A_CANCELLED	i
6	A_DECLINED	j
7	O_SELECTED O_CREATED O_SENT	l
8	O_SENT_BACK	n
9	O_ACCEPTED	o
10	O_CANCELLED	p
11	O_DECLINED	q
12	W_Beoordelen fraude	w

جدول ۱۶ - مسیرهای منحصر به فرد فرآیند بعد از ادغام و حذف فعالیت‌ها

Case_ID	Variant
173688	acdlengo
173691	acdelpnog
173694	acdelpnogx
173697	aj
173700	aj
…	…

برای کشف فرآیند بر اساس گزارش رویداد جدیدی که ساخته شده است، تنها کافی است ماتریس‌های *Frequency* و *Dependency* مطابق جداول ۱۷ و ۱۸، تشکیل شده و در نهایت گراف فرآیندی ارائه شود.

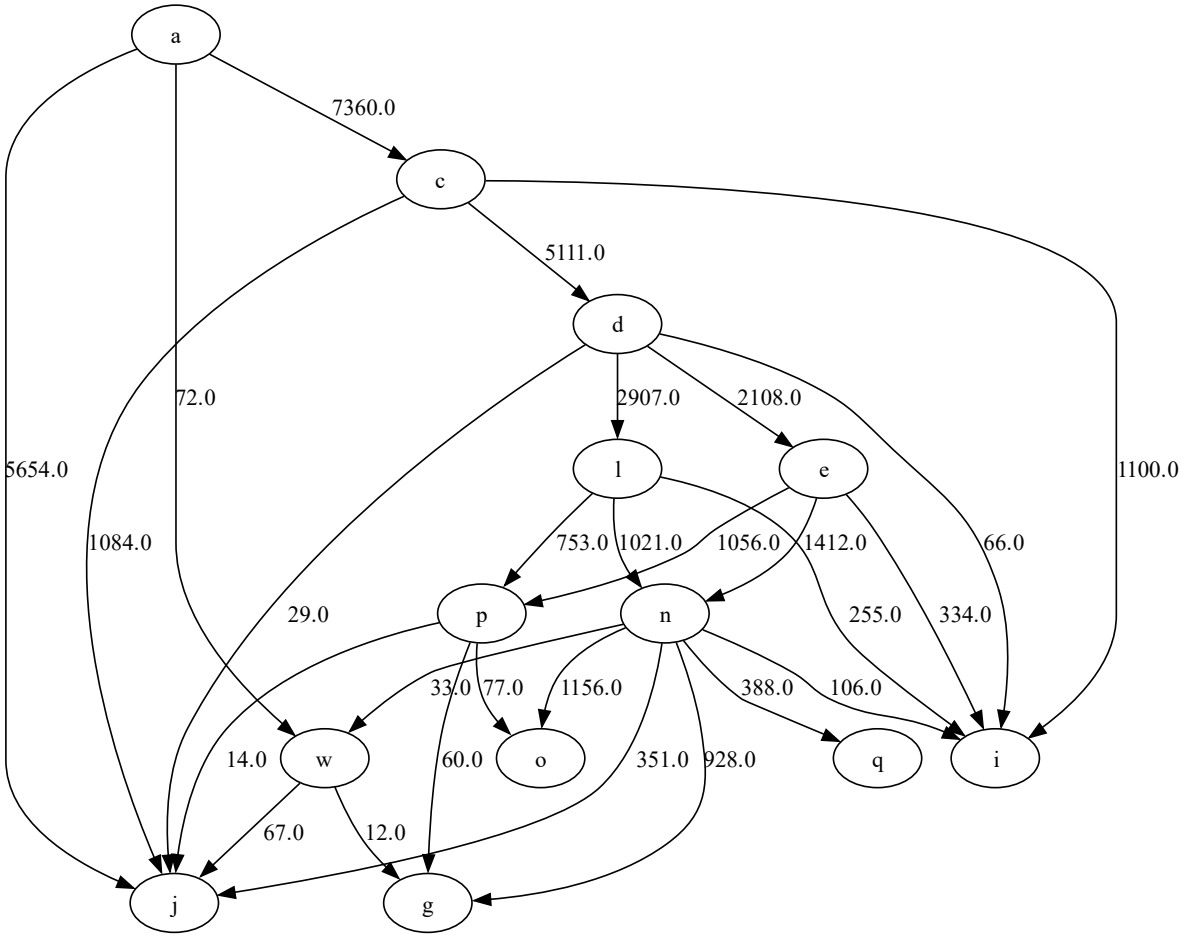
جدول ۱۷ – فراوانی روابط مستقیم بین فعالیت‌ها در گزارش رویداد فرآیند

$\|> L\|$	a	c	d	e	g	l	n	o	p	q	w	i	j
a	0	7360	0	0	0	0	0	0	0	0	72	1	5654
c	0	0	5111	0	0	0	0	0	0	0	3	1100	1084
d	0	0	0	2108	0	2907	0	0	0	0	0	66	29
e	0	0	0	0	0	2108	1412	0	1056	10	0	334	10
g	0	0	0	0	0	0	0	998	0	0	0	0	0
l	0	0	0	2907	0	0	1021	0	753	6	0	255	7
n	0	0	0	0	928	0	0	1156	207	388	33	106	351
o	0	0	0	0	1245	0	0	0	0	0	0	0	0
p	0	0	0	0	60	0	820	77	0	10	0	945	14
q	0	0	0	0	0	0	0	0	0	0	0	0	419
w	0	7	2	0	12	0	0	11	4	5	0	0	67
i	0	0	0	0	0	0	0	0	640	0	0	0	0
j	0	0	0	0	0	0	0	0	0	383	0	0	0

جدول ۱۸ – وابستگی بین فعالیت‌ها در گزارش رویداد فرآیند

$\|\rightarrow L\|$	a	c	d	e	g	l	n	o	p	q	w	i	j
a	0	1	0	0	0	0	0	0	0	0	0.986	0.5	1
c	-1	0	1	0	0	0	0	0	0	0	-0.364	0.999	0.999
d	0	-1	0	1	0	1	0	0	0	0	-0.667	0.985	0.967
e	0	0	-1	0	0	-0.159	0.999	0	0.999	0.909	0	0.997	0.909
g	0	0	0	0	0	0	-0.999	-0.11	-0.984	0	-0.923	0	0
l	0	0	-1	0.159	0	0	0.999	0	0.999	0.857	0	0.996	0.875
n	0	0	0	-0.999	0.999	-0.999	0	0.999	-0.596	0.997	0.971	0.991	0.997
o	0	0	0	0	0.11	0	-0.999	0	-0.987	0	-0.917	0	0
p	0	0	0	-0.999	0.984	-0.999	0.596	0.987	0	0.909	-0.8	0.192	0.933
q	0	0	0	-0.909	0	-0.857	-0.997	0	-0.909	0	-0.833	0	0.045
w	-0.986	0.364	0.667	0	0.923	0	-0.971	0.917	0.8	0.833	0	0	0.985
i	-0.5	-0.999	-0.985	-0.997	0	-0.996	-0.991	0	-0.192	0	0	0	0
j	-1	-0.999	-0.967	-0.909	0	-0.875	-0.997	0	-0.933	-0.045	-0.985	0	0

با در نظر گرفتن یک آستانه [1]با مقدار ۰.۹۲ گراف حاصل از طی شدن الگوریتم کشف فرآیند *Heuristics Miner* مطابق شکل ۴ قابل ارائه خواهد بود. این گراف که شماتیکی از یک درخت تصمیم‌گیری را نشان می‌دهد، کلیه روابط و حالات به همراه تعداد رخدادشان را شامل می‌گردد.



شکل ۴ - درخت تصمیم‌گیری حالت‌های مختلف درخواست وام

[1]. Threshold

۲- وضعیت درخواست‌های وام بانکی

برای بررسی میزان درخواست‌های وام یا اضافه برداشت در چهار وضعیت رد، کنسل، پذیرفته و یا تصمیم‌گیری نشده، کافی است پنج فعالیت زیر را که مشخص‌کننده شرح وضعیت هر مسیر منحصر به فرد در گزارش رویداد است را مطابق جدول ۱۹ در نظر داشت.

جدول ۱۹ - فعالیت‌های معرف وضعیت درخواست

No.	Activity	Letter
1	A_CANCELLED	i
2	A_DECLINED	j
3	O_ACCEPTED	o
4	O_CANCELLED	p
5	O_DECLINED	q

حال بر اساس این‌که کدام‌یک از فعالیت‌های فوق در انتهای یک مسیر رخ داده است، وضعیت کلیه‌ی درخواست‌های به ثبت رسیده مشخص می‌شود. رویکردی که در این زمینه پش گرفته شده است به شرح زیر می‌باشد؛

- اگر در یک مسیر فعالیت o وجود داشته باشد، مسیر دقیقاً از توسط همین فعالیت به دو قسمت تقسیم می‌گردد. سپس اگر در قسمت انتهایی مسیر (بخش دوم) سایر فعالیت‌های مرتبط با لغو یا رد شدن صورت نگیرد، آن درخواست پذیرفته شده محسوب می‌گردد.
- اگر در یک مسیر فعالیت j و q وجود داشته باشد، می‌توان آن‌ها را معرف گروه درخواست‌های رد شده دانست. اگر در یک مسیر j وجود داشته باشد ولی q خیر، آن‌گاه مسیر توسط فعالیت j به دو بخش تقسیم شده و نباید در ادامه‌ی مسیر سایر فعالیت‌های i، p و o حضور داشته باشند. اگر در یک مسیر q وجود داشته باشد ولی j خیر، آن‌گاه مسیر توسط فعالیت q به دو بخش تقسیم شده و نباید در ادامه‌ی مسیر سایر فعالیت‌های i، p و o حضور داشته باشند. اگر در یک مسیر هر دو فعالیت j و q وجود داشته باشد تنها کافی است هر یک را که دیرتر در مسیر ظاهر می‌شود را مبنای تعیین وضعیت قرار داد و مشابه آن‌چه گفته شد، پیش رفت.

- اگر در یک مسیر فعالیت i و p وجود داشته باشد، می‌توان آن‌ها را معرف گروه درخواست‌های لغو شده دانست. اگر در یک مسیر i وجود داشته باشد ولی p خیر، آن‌گاه مسیر توسط فعالیت i به دو بخش تقسیم شده و نباید در ادامه‌ی مسیر سایر فعالیت‌های j، q و o حضور داشته باشند. اگر در یک مسیر p وجود داشته باشد ولی i خیر، آن‌گاه مسیر توسط فعالیت p به دو بخش تقسیم شده و نباید در ادامه‌ی مسیر سایر فعالیت‌های j، q و o حضور داشته باشند. اگر در یک مسیر هر دو فعالیت i و p وجود داشته باشد تنها کافی است هر یک را که دیرتر در مسیر ظاهر می‌شود را مبنای تعیین وضعیت قرار داد و مشابه آن‌چه گفته شد، پیش رفت.
- اگر هیچ یک حالات بالا برای یک درخواست مشخص در فرآیند صدق نکند، درخواست به وضعیت تصمیم‌گیری نشده در خواهد آمد.

جدول ۲۰، نشان‌دهنده وضعیت درخواست‌های به ثبت رسیده است. همان‌طور که مشهود است، اگر یک درخواست به حالت پذیرفته شده در بیاید، دیگر امکان ندارد حالاتی نظیر رد، لغو و یا تصمیم‌گیری نشده برای‌ش صدق کند. بنابراین تنها یکی از حالات ممکن قابل رخ دادن می‌باشد.

جدول ۲۰ – وضعیت درخواست‌های ثبت شده در فرآیند

Case_ID	Variant	Start Event	End Event	Amount_Request	Accepted	Declined	Cancelled	No Decision
173688	Abcssdkelmtsttttnutugfohu	a	u	20000	TRUE	FALSE	FALSE	FALSE
173691	abcssssdeklmtstttkplmtttttnutuuuuuofghu	a	u	5000	TRUE	FALSE	FALSE	FALSE
173694	abcsssssssdeklmtstkplmttttttttkplmttttttttt...	a	x	7000	TRUE	FALSE	FALSE	FALSE
173697	abj	a	j	15000	FALSE	TRUE	FALSE	FALSE
173700	abj	a	j	5000	FALSE	TRUE	FALSE	FALSE
...	...	...	...	...	...	...	...	...

با بررسی کلیه درخواست‌های موجود، می‌توان متوجه شد متوسط مقدار وام‌های پذیرفته، رد، لغو و یا تصمیم‌گیری نشده در فرآیند به چه صورت است. جدول ۲۱، اطلاعات کلی در این رابطه را ارائه می‌دهد.

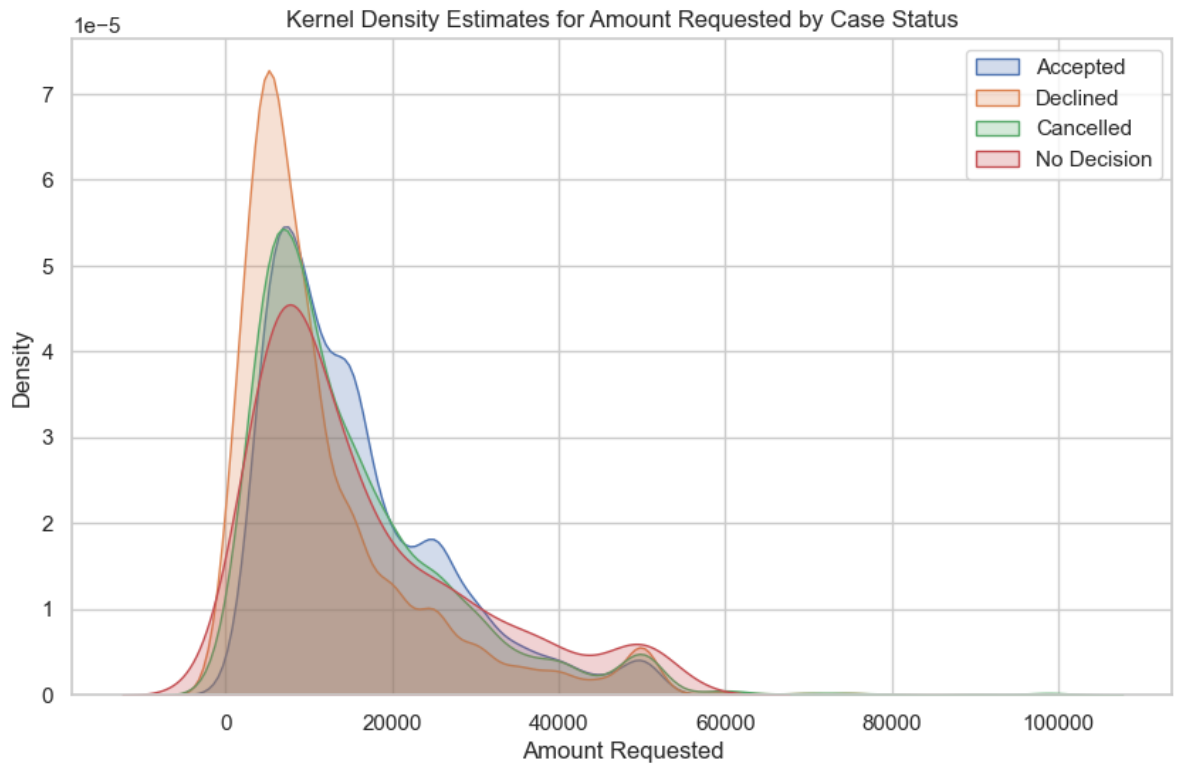
جدول ۲۱ - میزان وام دریافتی در هر حالت تصمیم‌گیری

Status	Min	Mean	Median	Max
Accepted	1000	15705.46	13000	99000
Declined	1	12191.03	8000	99999
Cancelled	0	15280.78	10000	99999
No Decision	300	16314.88	10000	55000

برای بررسی این‌که هر یک از وضعیت‌های فوق در چه صورتی رخ می‌دهند، می‌توان ادله‌ی زیر را ارائه نمود. به عنوان مثال، اگر در یک فرآیند درخواست ثبت شده مشکلی از منظر اطلاعات و مبلغ درخواستی نداشته باشد و مشتری از عهده‌ی بازپرداخت آن بربیاید، درخواست پذیرفته می‌شود، ولیکن درخواست ثبت شده می‌تواند از سمت مشتری لغو شده و یا از طرف بانک به علت فقدان اطلاعات یا واجد شرایط نبودن اهدای وام، رد گردد. همچنین گاهاً فرآیند با مشخص شدن موارد مشکوک به کلاهبردی خاتمه یافته است، که در این شرایط درخواست به وضعیت عدم تصمیم‌گیری در خواهد آمد.

۲-۱- علل موثر در پذیرش یا عدم پذیرش درخواست‌ها

توقع می‌رود تنها متغیری که بر تعیین شدن وضعیت یک درخواست تأثیر خواهد داشت، مقدار وام درخواست شده از سمت مشتری باشد. ولیکن شکل ۵ ارائه شده است تا نشان دهد، حتی میزان وام درخواست شده از سمت مشتری نیز هیچ تأثیر چشمگیر و یا رابطه‌ی معناداری بر مشخص شدن وضعیت یک درخواست (لغو، رد، پذیرش و عدم تصمیم‌گیری) ندارد.



شکل ۵ – توزیع مقدار وام بر اساس وضعیت‌های ممکن درخواست

۳- تحلیل فرآیند

در این بخش، فرآیند از دو منظر مدت زمان اجرا و بهره‌وری منابع به کار گرفته شده، بررسی می‌گردد.

۳-۱- مدت زمان اجرای فرآیند

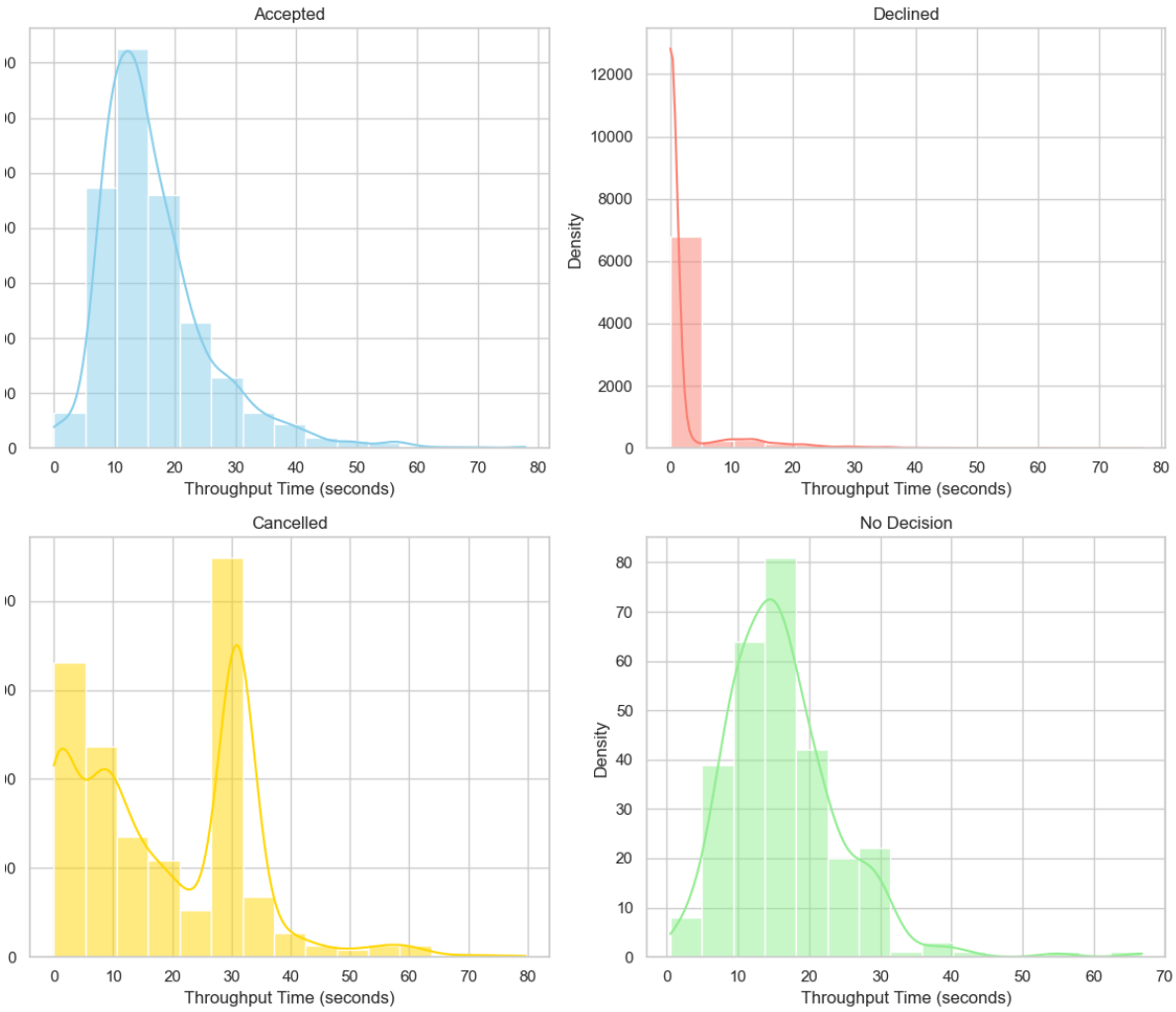
محاسبه‌ی مدت زمان لازم برای اجرای هر درخواست، یا به عبارتی هر تکرار از فرآیند، بر اساس مقاطع زمانی گزارش رویداد میسر می‌گردد. تنها کافی است برای هر درخواست، حد فاصل میان زمان خاتمه‌ی آخرین فعالیت و زمان ثبت درخواست ارزیابی گردد. بر این اساس هدف فوق تحقق می‌یابد. جدول ۲۲، مدت زمان محاسبه شده برای هر درخواست را نشان می‌دهد.

جدول ۲۲ - مدت زمان رسیدگی به هر درخواست

Case_ID	Variant	Amount_Request	Accepted	Declined	Cancelled	No Decision	Throughput Time
173688	abcssdkelmtsttttnutugfohu	20000	TRUE	FALSE	FALSE	FALSE	12 days 09:58:52.480000
173691	abcssssdeklmtstttkplmtttttnutuuuuuofghu	5000	TRUE	FALSE	FALSE	FALSE	9 days 06:08:36.377000
173694	abcsssssssdeklmtstkplmttttttttkplmttttttttt...	7000	TRUE	FALSE	FALSE	FALSE	137 days 04:18:56.012000
173697	abj	15000	FALSE	TRUE	FALSE	FALSE	0 days 00:00:37.555000
173700	abj	5000	FALSE	TRUE	FALSE	FALSE	0 days 00:00:41.143000
...	...	...	...	...	...	...	...
214364	abcssssdeklmtstkplmttttttttnut	5000	FALSE	FALSE	TRUE	FALSE	8 days 11:39:23.786000
214367	abj	500	FALSE	TRUE	FALSE	FALSE	0 days 00:00:40.860000
214370	abrrjr	20000	FALSE	TRUE	FALSE	FALSE	0 days 09:59:25.879000
214373	abrrcsrsdkelmtstt	8500	FALSE	FALSE	FALSE	TRUE	9 days 13:07:45.115000
214376	abrrjr	15000	FALSE	TRUE	FALSE	FALSE	0 days 09:36:24.526000

به صورت کلی، میانگین مدت زمان اجرای فرآیند برابر است با ← *8 days 14:55:16.341417*

شکل ۶، هیستوگرام توزیع مدت زمان اجرای درخواست‌ها را بر اساس وضعیت درخواست (پذیرش، لغو، رد و یا عدم تصمیم‌گیری) نشان می‌دهد.



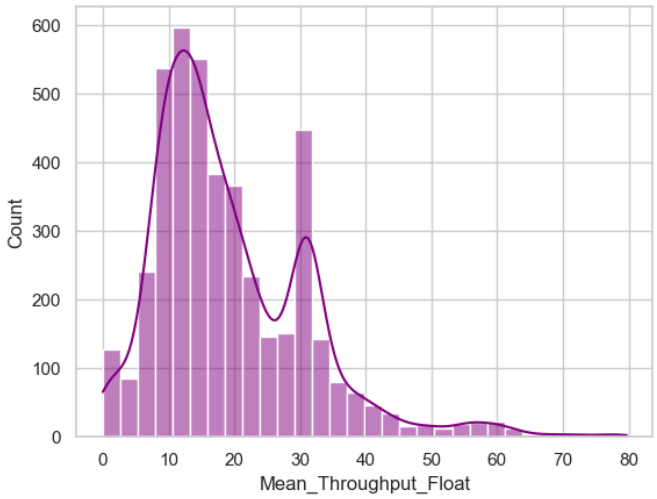
شکل ۶ – هیستوگرام توزیع مدت زمان اجرای درخواست‌ها بر اساس وضعیت درخواست

برای درک بهتر مدت زمان اجرای فرآیند، از رویکرد دیگری نیز استفاده شده است که به شرح زیر است. مطابق جدول ۲۳، می‌توان کلیه مسیرهای منحصر به فرد فرآیند را شناسایی نموده و سپس میانگین مدت زمان اجرای هر مسیر منحصر به فرد را محاسبه نمود. بنابر این رویکرد، شکل ۷ نتیجه می‌شود که به علت داشتن دو مد[1] در هیستوگرام توزیع خود، به دو نمودار مجزا همراه با یک مد[2]، مطابق شکل ۸ تجزیه می‌گردد.

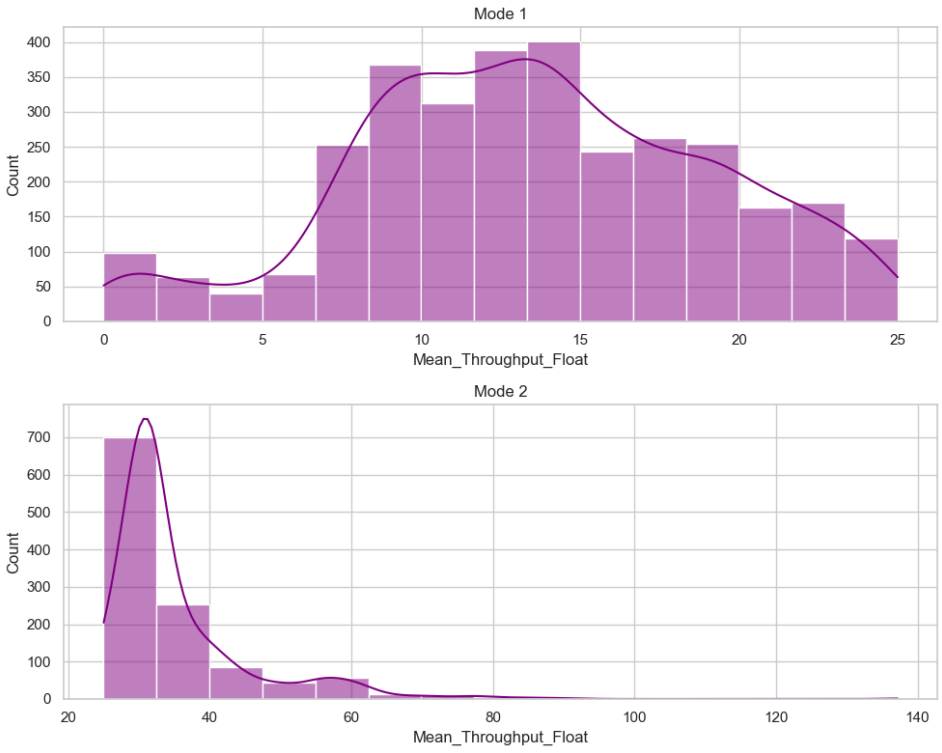
[1]. Bimodal
[2]. Unimodal

جدول ۲۳ – متوسط مدت زمان هر مسیر منحصر به فرد فرآیند

No.	Variant	Iterations	Mean_Throughput_Float
0	abj	3429	0.000446
1	abrrjr	1872	0.215596
2	abrrrrjr	271	0.288694
3	abrrcsrsjs	209	0.324235
4	abcssjs	160	0.262061
…	…	…	…



شکل ۷ – هیستوگرام توزیع مدت زمان اجرای مسیرهای فرآیند همراه با دو مد



شکل ۸ – هیستوگرام توزیع مدت زمان اجرای مسیرهای فرآیند همراه با یک مد

اطلاعات دیگری که در خصوص مدت زمان اجرای هر درخواست می‌توان بدست آورد، طبقه‌بندی براساس رویدادهای پایانی ممکن است. جدول ۲۴ نشان می‌دهد، به عنوان مثال، متوسط مدت زمان اجرای درخواست‌هایی که با رویداد x به پایان رسیده‌اند، ۵۱ روز می‌باشد. مسیرهایی با رویداد پایانی مذکور تنها چهار مرتبه در کل گزارش رویداد ظاهر شده است.

جدول ۲۴ - متوسط مدت زمان هر مسیر منحصر به فرد بر اساس رویدادهای پایانی

End Event	Min Duration	Mean Duration	Max Duration	Occurrences	Ratio
x	10 days 01:33:45.029000	51 days 15:56:08.999000	137 days 04:18:56.012000	4	0.03
p	2 days 18:51:53.401000	33 days 13:51:19.174577061	71 days 10:45:53.799000	279	2.13
i	0 days 00:48:25.214000	31 days 22:31:42.223638168	91 days 09:55:36.161000	655	5
v	0 days 05:48:32.664000	20 days 17:37:51.176376106	84 days 06:41:26.256000	452	3.45
u	0 days 00:11:48.376000	16 days 15:42:21.382077539	85 days 21:02:24.197000	2747	20.99
t	0 days 00:08:27.914000	16 days 04:52:09.690982170	82 days 18:34:27.246000	1290	9.86
g	12 days 20:59:32.871000	12 days 20:59:32.871000	12 days 20:59:32.871000	1	0.01
s	0 days 00:01:06.618000	2 days 21:21:10.724641567	32 days 15:17:09.137000	1939	14.82
w	0 days 00:52:08.010000	2 days 14:52:10.877333333	12 days 23:38:32.096000	57	0.44
r	0 days 00:01:17.481000	0 days 05:52:55.881317815	6 days 16:47:40.875000	2234	17.07
j	0 days 00:00:01.855000	0 days 00:00:38.524777486	0 days 00:01:59.934000	3429	26.2

۳-۲- بهره‌وری منابع فرآیند

در این بخش رویکردی ارائه می‌گردد تا بتوان بر اساس آن در مورد بهره‌ور بودن یا نبودن منابع موجود در فرآیند قضاوت نمود. یکی از مشکلات گزارش رویدادهای فراهم شده، فقدان مقاطع زمانی شروع هر فعالیت است. ثبت شدن مقاطع زمانی شروع رویدادها از آن جهت باعث سهولت کار می‌گشت که با در نظر گرفتن اختلاف زمان آغاز و پایان هر فعالیت، مدت زمان اجرای آن فعالیت قابل محاسبه می‌شد. ولیکن اکنون، رفع این چالش کمی دشوار خواهد بود. بدین جهت فعالیت‌های به کار رفته در هر مسیر منحصر به فرد را می‌توان به صورت زیر بررسی نمود [10]. همان‌طور که پیش‌تر ذکر شد، بسیاری از فعالیت‌های فرآیند به صورت خودکار و سیستماتیک صورت می‌گیرد، بنابراین می‌توان از مدت زمان انجام آن‌ها که چیزی در حدود ثانیه است، صرف نظر کرد. بنابراین از میان کلیه‌ی فعالیت‌های موجود در فرآیند فقط آن‌هایی در مدت زمان اجرای فرآیند تأثیرگذار هستند که *Transition* آن‌ها همواره دارای یک *Start* و *Complete* مشخص هستند. بررسی گزارش رویداد حاکی از آن است که تنها فعالیت‌های زیر دارای چنین مقاطعی می‌باشند: $\{u, t, s, r, v, w\}$

لازم به ذکر است ۲ مورد از رکوردها تنها دارای مقاطع زمانی شروع هستند و به پایان نرسیده‌اند که از مجموعه‌ی داده‌ها حذف شدند، همچنین ۱۰۳۱ مورد از رکوردها نیز دارای چندین مقاطع پایانی بودند که برای حفظ منطق محاسبات، آخرین مقطع زمانی آن‌ها در نظر گرفته شده است. جدول ۲۵، مشخص‌کننده‌ی طولانی‌ترین فعالیت هر درخواست انجام شده در فرآیند می‌باشد.

جدول ۲۵ - ردیابی طولانی‌ترین فعالیت‌های هر درخواست

Case_ID	Activity	Maximum Duration	Maximum Duration Float	Resource
173688	u	0 days 00:32:10.101000	32.16835	10629
173691	u	0 days 00:18:18.281000	18.30468	10809
173694	u	0 days 00:40:39.578000	40.65963	10609
173703	s	0 days 00:21:45.599000	21.75998	10912
173706	s	0 days 00:09:36.365000	9.606083	112
...	...	...	...	...

قابل توجه است که یک درخواست می‌تواند چندین فعالیت از مجموعه‌ی مذکور را دارا باشد، ولیکن برای پیشبرد محاسبات تنها فعالیتی لحاظ می‌شود که بیشترین مدت زمان انجام را داشته باشد. از آن‌جایی که به دنبال منابعی هستیم که بیشترین اتلاف زمانی را داشته‌اند، بیشترین مدت انجام هر فعالیت در درخواست ثبت شده مد نظر است.

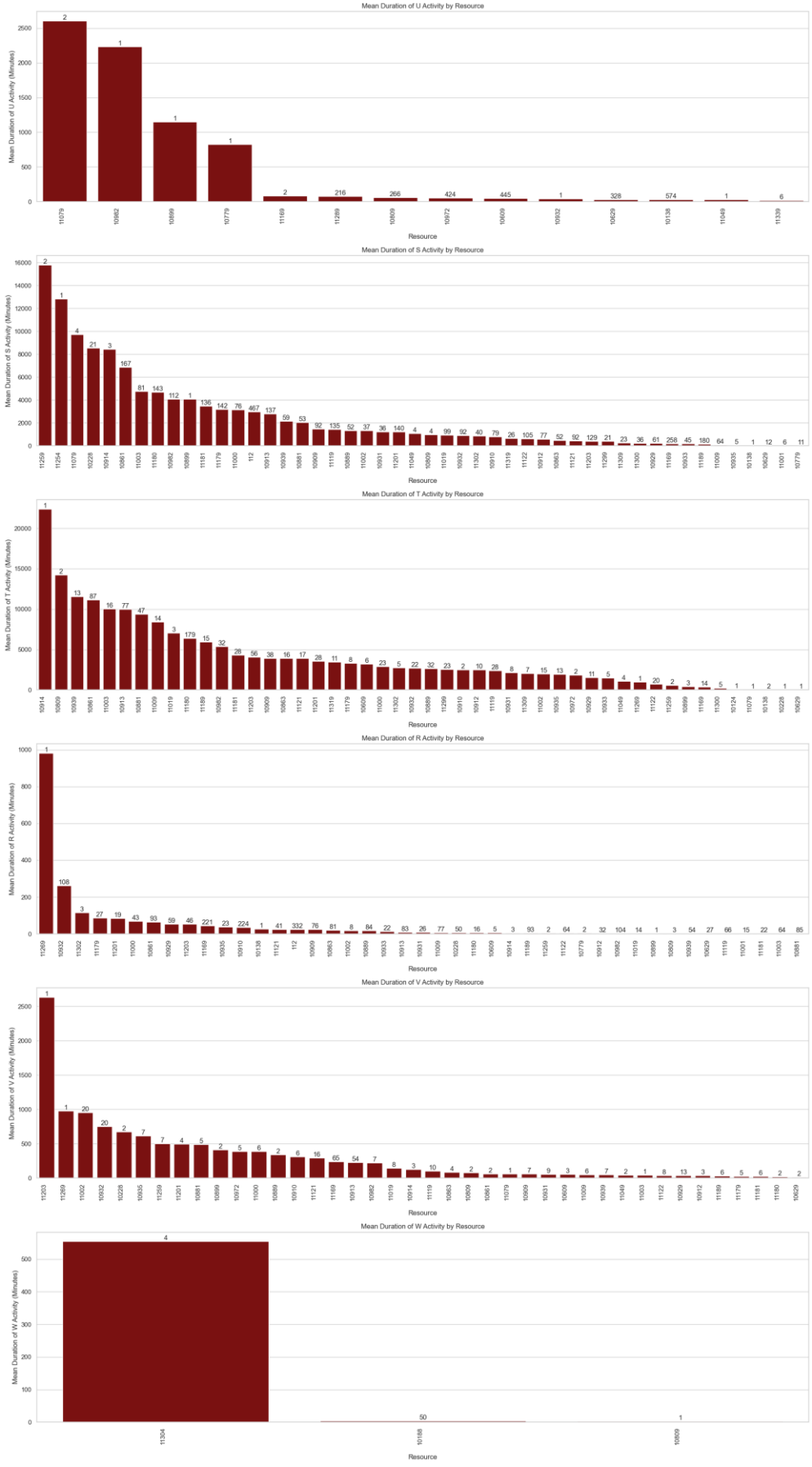
همچنین جدول ۲۵ نشان می‌دهد که در هر درخواست، یکی از شش فعالیت مذکور بیشنیه زمان انجام را دارد و هر یک از این فعالیت‌ها توسط منابع مختلفی انجام می‌گردد، بنابراین شکل ۹، مشخص می‌کند که هر فعالیت توسط چه منابعی انجام گرفته است. علاوه بر این در این نمودار می‌توان متوسط مدت زمان انجام هر فعالیت را توسط هر منبع نیز مشاهده نمود.

برای تحلیل شکل ذیل، به عنوان نمونه می‌توان فعالیت u را مثال زد که در آن سه منبع ۱۱۰۷۹، ۱۰۹۸۲ و ۱۰۸۹۹ به ترتیب بیشترین مدت زمان اجرا و کمترین تکرار را در بین منابع فعال دارند.

جدول ۲۶ نیز اطلاعاتی در خصوص بیشنیه‌ی مدت زمان انجام شش فعالیت مذکور ارائه می‌دهد.

جدول ۲۶ - بیشینه مدت زمان اجرای فعالیت‌های زمان‌دار

Activity	Maximum Duration Float
r	35.836227
s	2054.772755
t	5678.005224
u	50.696471
v	293.149063
w	43.937544



شکل ۹ – نمودار بهره‌وری منابع به کار گرفته شده در فرآیند

۴- فرآیندکاوی اولیه

در این بخش با استفاده از الگوریتم‌های مختلف، مدل فرآیندی بر اساس گزارش رویداد اصلی، بدون هیچ‌گونه تغییری کشف شده و سپس شاخص‌های انطباق آن مورد بررسی و تحلیل واقع می‌گردد.

۴-۱- کشف فرآیند

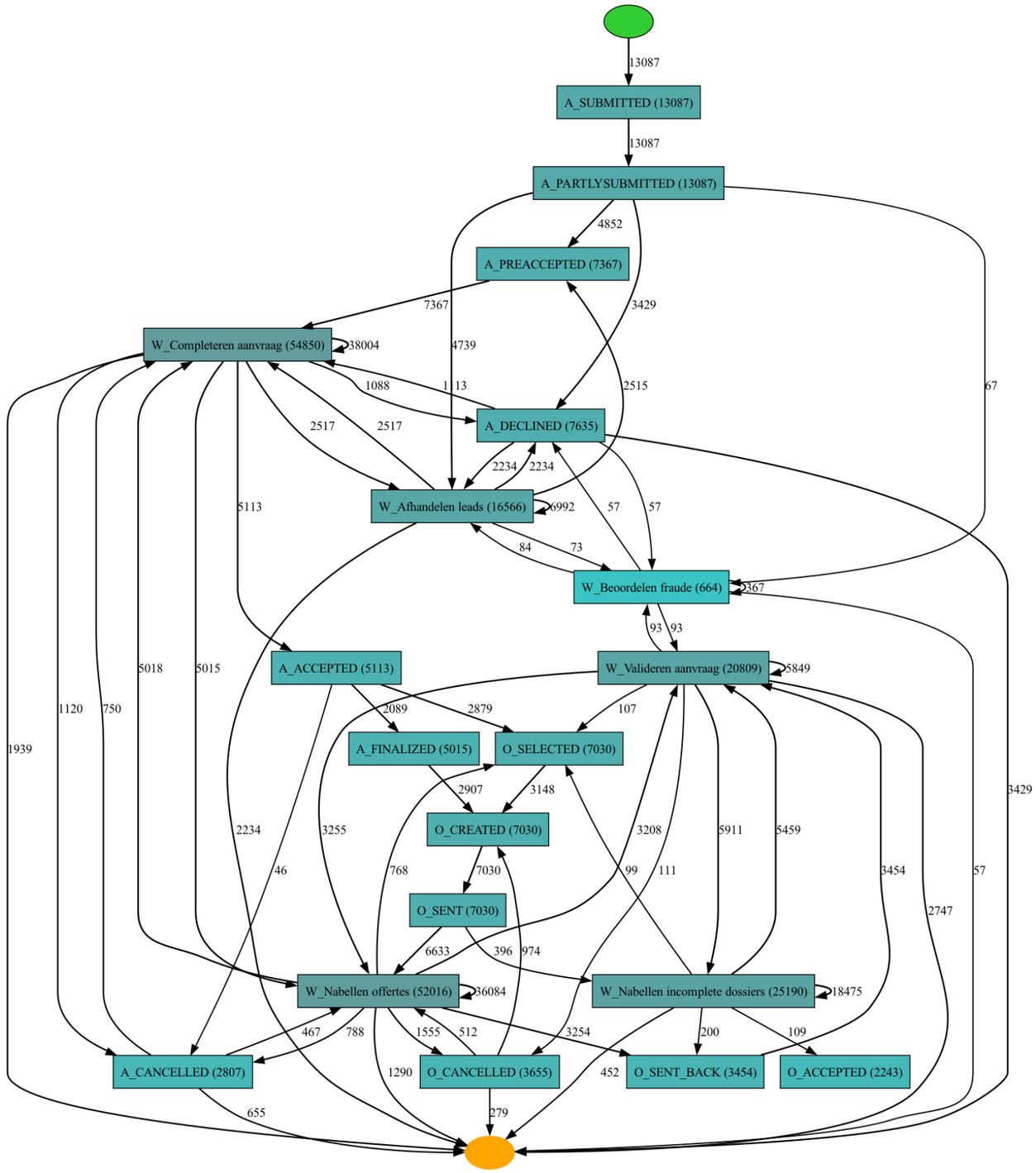
برای درک بهتر آن‌چه در حقیقت امر رخ داده است، مدل‌های فرآیندی متفاوتی بر روی گزارش رویداد اعمال و نتایج آن‌ها ارائه می‌شود. انواع مدل‌های بررسی شده به شرح زیر است:

- *Heuristics Miner – DFG*

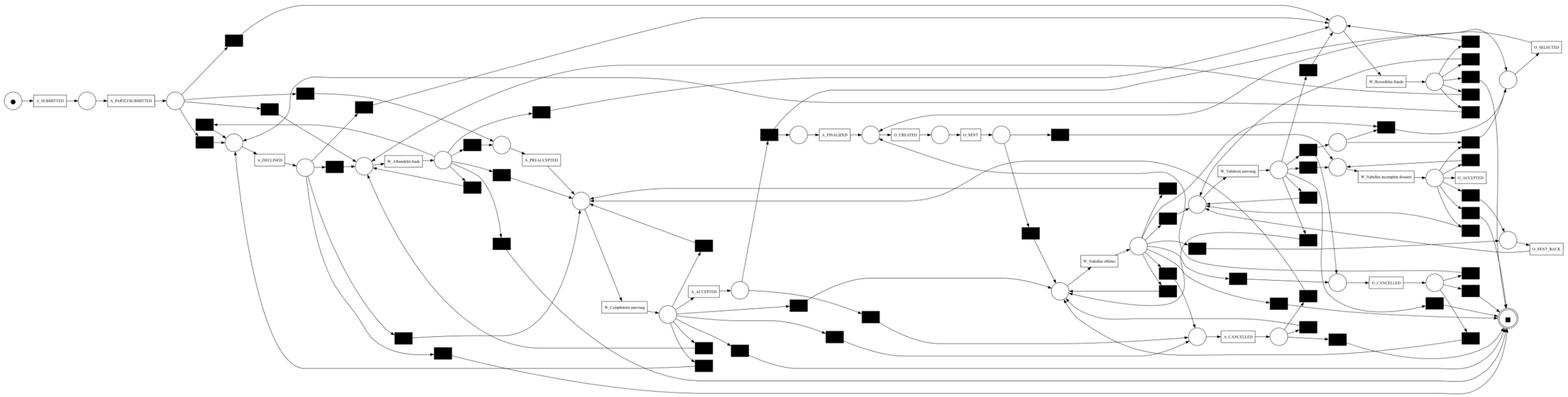
این مدل با در نظر گرفتن آستانه‌های زیر بدست می‌آید.

Min DFG occurrence → 40
DFG Cleaning Noise Threshold → 0.02
Dependency Threshold → 0.92
AND Measure Threshold → 0.95

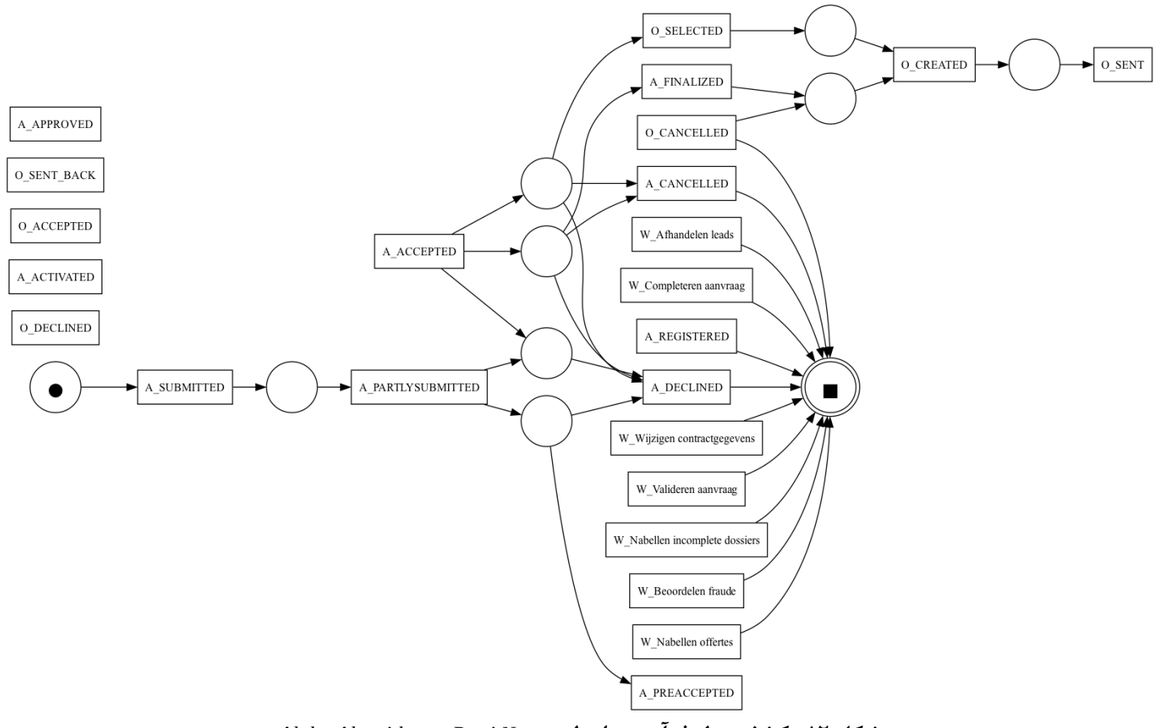
- *Heuristics Miner – Petri Net*
- *Alpha Algorithm – Petri Net*
- *Petri Net Heuristics*
- *Inductive Miner – BPMN*



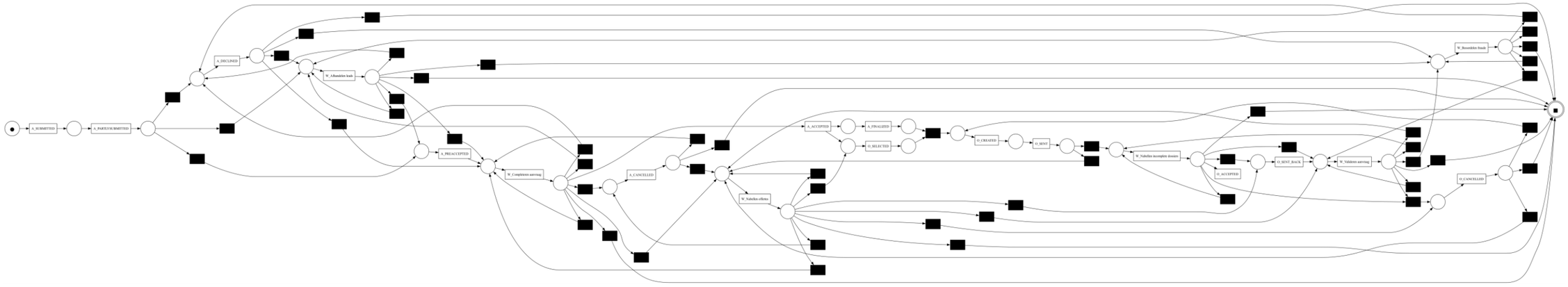
شکل ۱۰ - کشف مدل فرآیندی از طریق الگوریتم *Heuristics Miner*



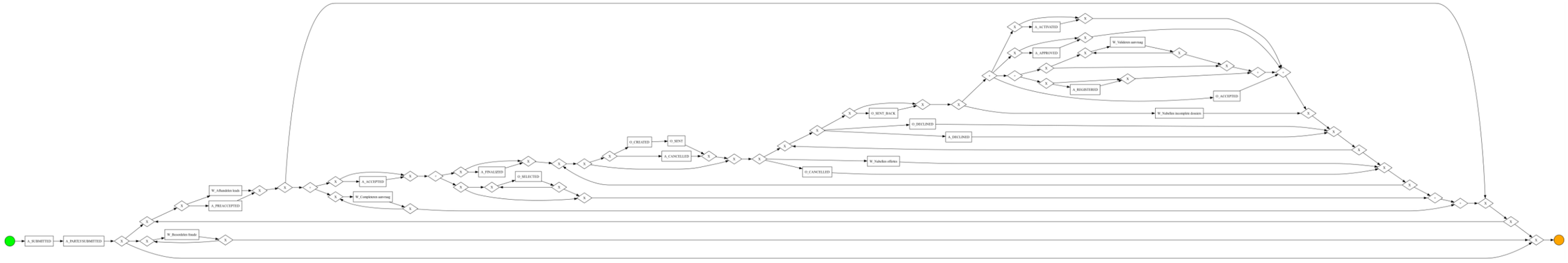
شکل ۱۱ - کشف مدل فرآیندی از طریق الگوریتم *Heuristics Miner – Petri Net*



شکل ۱۲ - کشف مدل فرآیندی از طریق *Alpha Algorithm — Petri Net*



شکل ۱۳ - کشف مدل فرآیندی از طریق *Petri Net Heuristics*



شکل ۱۴ - کشف مدل فرآیندی از طریق *Inductive Miner – BPMN*

۴-۲- بررسی انطباق فرآیند

بررسی انطباق تکنیکی است برای مقایسه یک مدل فرآیندی استخراج شده با گزارش رویداد همان فرآیند. هدف بررسی این است که آیا گزارش رویداد با مدل به دست آمده مطابقت دارد یا خیر و بالعکس. تکنیک‌های بسیاری برای بررسی انطباق وجود دارد. دو رویکردی که اغلب مورد استفاده قرار می‌گیرند، به شرح زیر هستند [11]:

- ***Token-Based Replay***
- ***Alignments***

رویکرد دوم مبنای تحلیل فرآیند درخواست وام بانکی خواهد بود. همچنین لازم به ذکر است، برای مقایسه رفتار قابل مشاهده در مدل استخراج شده با گزارش رویداد موجود، چهار شاخص کیفی در نظر گرفته می‌شود که در ادامه به توضیح آن پرداخته می‌شود [11]:

- برازش[1]: مدل کشف شده باید رفتاری را که در گزارش رویداد وجود دارد، نشان دهد یا به عبارتی قادر به بازتولید رفتاری باشد که در گزارش رویداد قابل ردیابی است. لازم به ذکر است اگر این شاخص مقدار یک را به خود بگیرد، نشانه‌ی افراط در برازش است و آن‌چنان ایده‌آل نیست. چرا که اگر مدل کشف شده فقط قابلیت بازیابی مسیرهای موجود را داشته باشد، شاخص تعمیم‌پذیری را نادیده خواهد گرفت.
- دقت[2]: شاخص دقت اندازه‌گیری می‌کند که تا چه مقدار رفتارهای مجاز توسط مدل کشف شده است که در گزارش رویداد نیز قابل مشاهده است. مدلی که اجازه‌ی شکل‌گیری رفتارهایی را می‌دهد که در گزارش رویداد وجود ندارد، نادقیق در نظر گرفته می‌شود. مقدار دقت یک، ممکن است بسیار محدودکننده بوده و اجازه‌ی هیچ‌گونه انحرافی را از گزارش رویداد نخواهد داد. بنابراین هرچند دقت بالا مطلوب است، ولیکن مقدار یک همیشه ایده‌آل نیست.

---

[1]. Fitness
[2]. Percision

- تعمیم‌پذیری[1]: این معیار مربوط به کمی کردن میزان تعمیم رفتار یک مدل به رفتاری است که در گزارش رویداد مشاهده نشده است، ولیکن احتمال رخداد آن وجود دارد. به تعبیری دیگر، مدل استخراج شده باید قادر باشد، مسیرهای بیشتری را جز آن‌چه اتفاق افتاده است، پیش‌بینی نماید. افزایش میزان تعمیم‌پذیری (شناسایی رفتارهای ناشناخته)، شاخص دقت را در بررسی انطباق به خطر می‌اندازد.
- سادگی[2]: هر چه مدل استخراج شده ساده‌تر باشد، مطلوب‌تر است، چرا که باعث تسهیل امر مدیریت و درک فرآیند می‌گردد. هر چه این شاخص مقدار کمتری به خود گیرد، بهتر است، چرا که دلالت بر کاهش پیچیدگی مدل فرآیندی دارد.

در ادامه، کلیه‌ی شاخص‌ها نامبرده برای الگوریتم‌های به کار رفته در قالب جدول ۲۷ ارائه شده است:

جدول ۲۷- محاسبه‌ی شاخص‌های انطباق فرآیند کشف شده

Algorithm	Fitness	Percision	Generalization	Simplicity
Heuristics Miner	0.95	0.43	0.95	0.54
Alpha Algorithm	-	0.97	-	0.94
Petri Net Heuristics	0.97	0.46	0.96	0.55

میزان بالا بودن شاخص اول، نشان از بیش‌برازش[3] مدل است. از آن‌جایی که در همه‌ی مدل‌های کشف‌شده شاخص تعمیم‌پذیری بسیار بالاست، دقت مدل کاهش یافته است. همچنین عمده‌ی مدل‌های کشف شده از معیار سادگی و قابلیت درک آسان نیز به دور بوده‌اند.

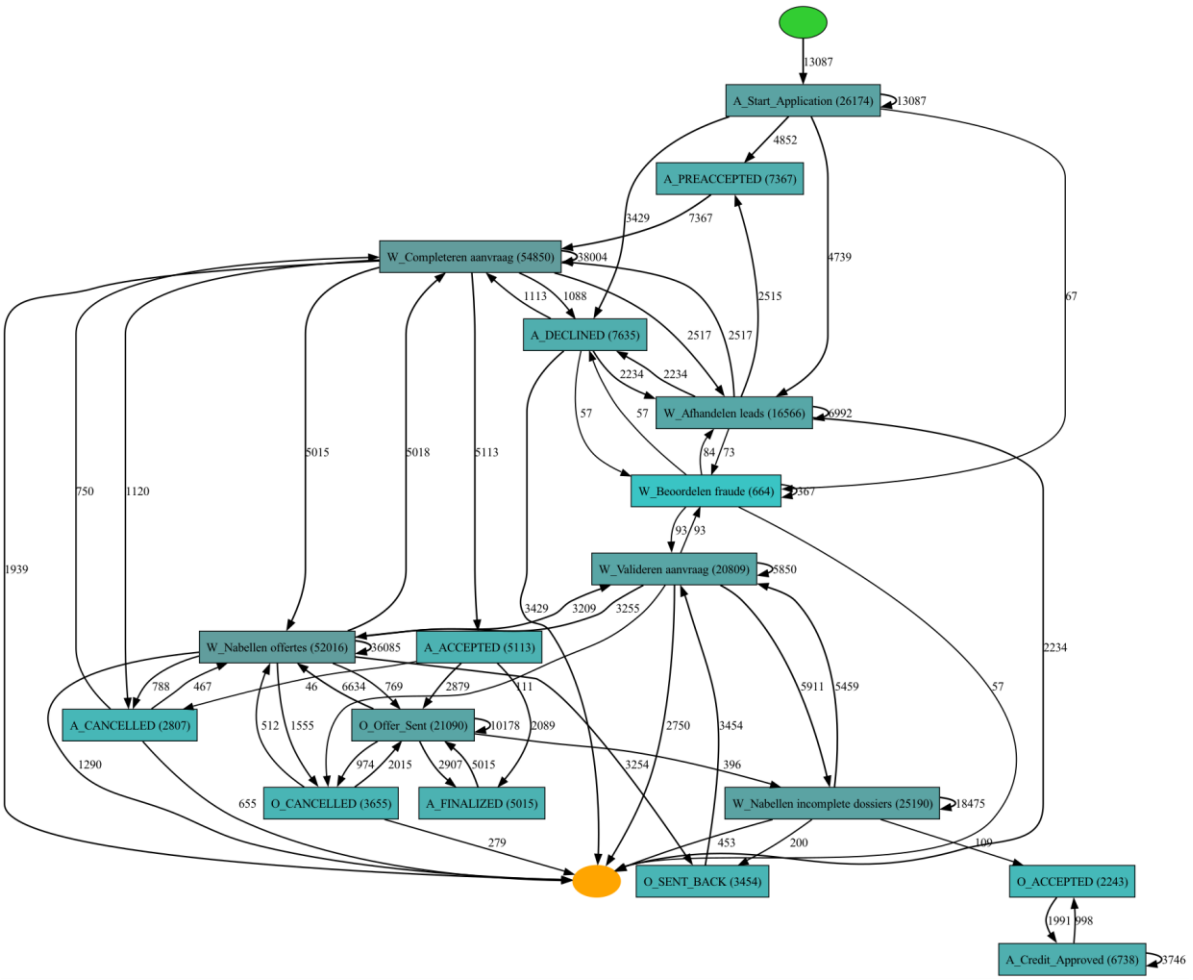
[1]. Generalization
[2]. Simplicity
[3]. Over Fitness

۴-۳- بهبود مدل فرآیند

در وهله‌ی آخر، پشنهاداتی در خصوص بهبود مدل فرآیندی کشف شده ارائه می‌گردد. اولاً که فعالیت {*W_Wijzigen contractgegevens*} در کل گزارش رویداد تنها ۰.۰۰۵ درصد ظاهر شده است، بنابراین این فعالیت را حذف نموده و ثانیاً ادغام فعالیت‌های زیر به دلایلی که پیشتر مطرح شد، صورت خواهد گرفت.

- A_SUBMITTED و A_PARTLYSUBMITTED ← A_Start_Application
- A_REGISTERED، A_APPROVED و A_ACTIVATED ← A_Credit_Approved
- O_CREATED، O_SELECTED و O_SENT ← O_Offer_Sent

در ادامه مدل کشف شده با شاخص برازش ۰.۹۶ مطابق شکل ۱۵ ارائه شده است.



شکل ۱۵ - کشف مدل فرآیندی از طریق الگوریتم *Heuristics Miner*

۵- فرآیندکاوی ثانویه

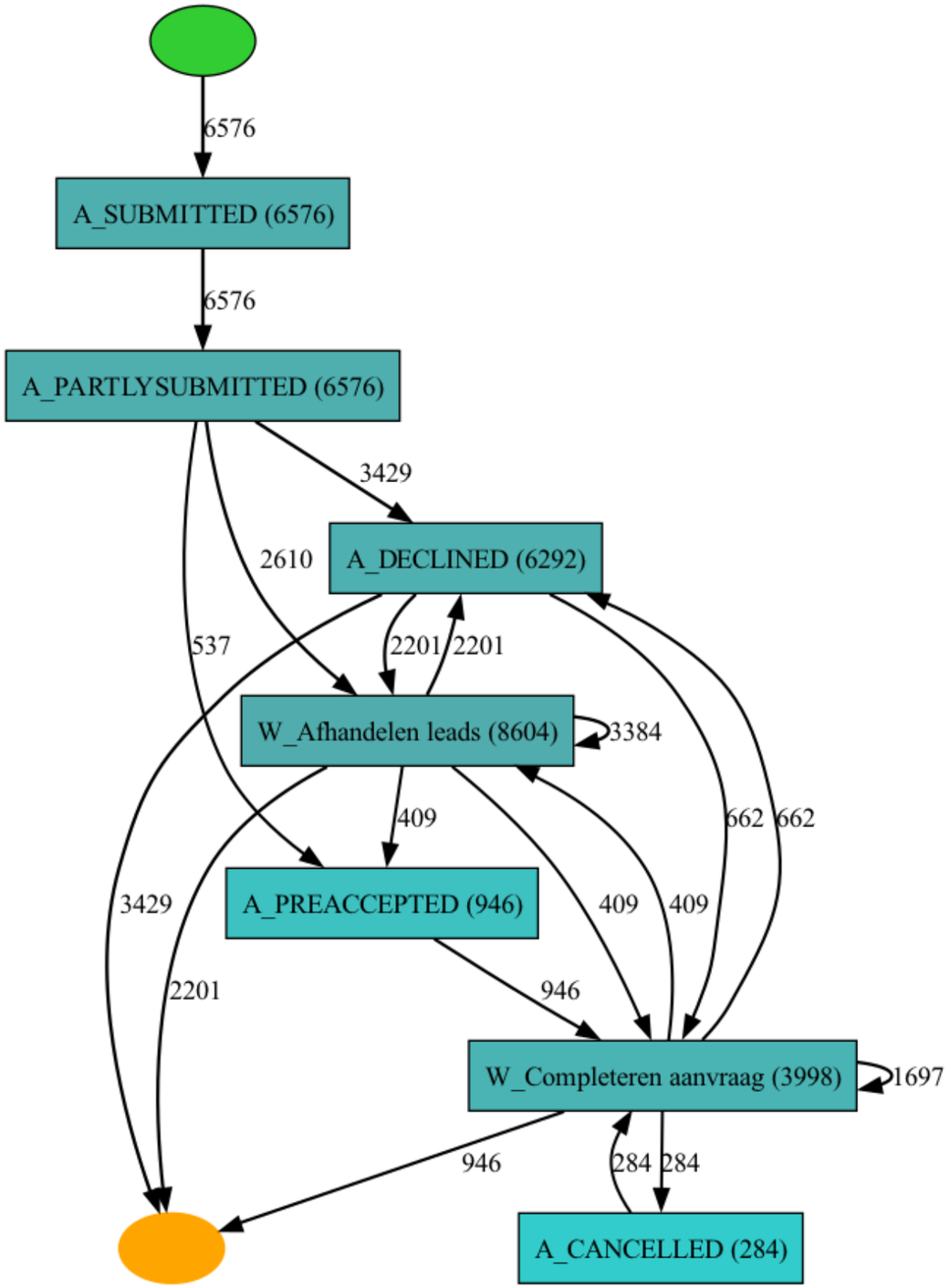
به عنوان رویکردی دیگر در راستای فرآیندکاوی گزارش رویداد اصلی، تصمیم گرفته شد که دسته‌ای از مسیرهای فرآیندی که حداقل ۵۰ درصد کل پرونده‌ها را در گزارش رویداد شامل می‌شود، در یک گزارش رویداد جدید ذخیره نموده و سپس کشف مدل بر روی لاگ‌های جدید صورت گیرد و مقادیر شاخص‌های انطباق محاسبه شود. با توجه به این فرضیه، ۱۲ مسیر اصلی واجد شرایط مطابق جدول ۲۸ ارائه شده است.

جدول ۲۸- مسیرهای منحصر به فرد پرتکرار در گزارش رویداد

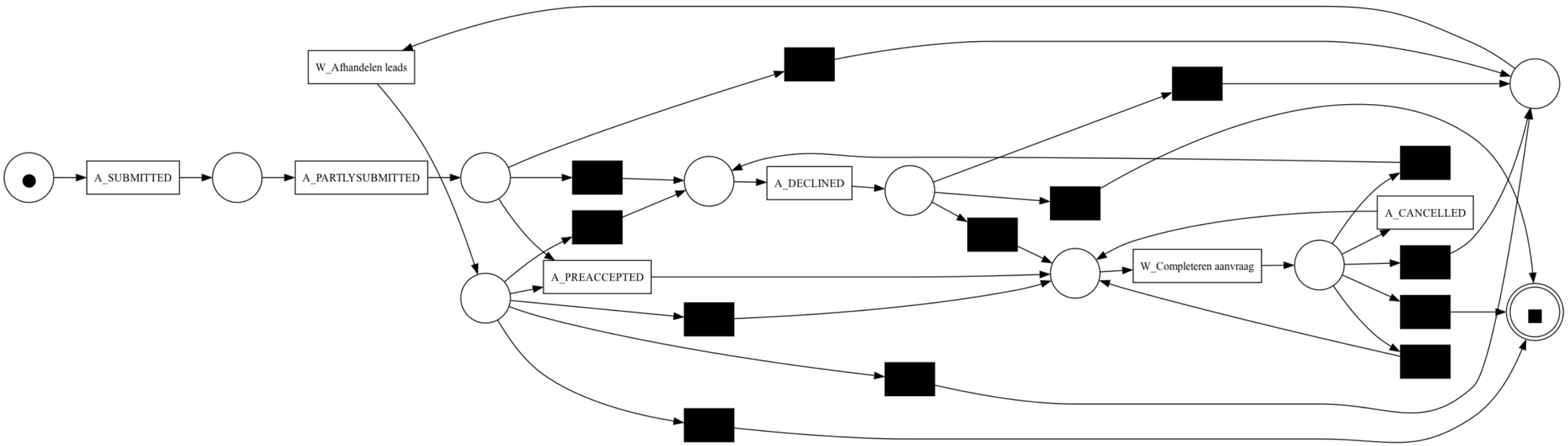
No.	Variant	Iterations	Cumulative Percentage
11	abrrrrrrjr	58	0.501948
10	abcssssssis	63	0.506762
9	abrrcsrssssjs	74	0.512417
8	abcssssis	87	0.519065
7	abcssssjs	93	0.526171
6	abrrcsrsssjs	126	0.535799
5	abcssis	134	0.546038
4	abcssjs	160	0.558264
3	abrrcsrsjs	209	0.574234
2	abrrrrjr	271	0.594942
1	abrrjr	1872	0.737984
0	abj	3429	1

۵-۱- کشف فرآیند

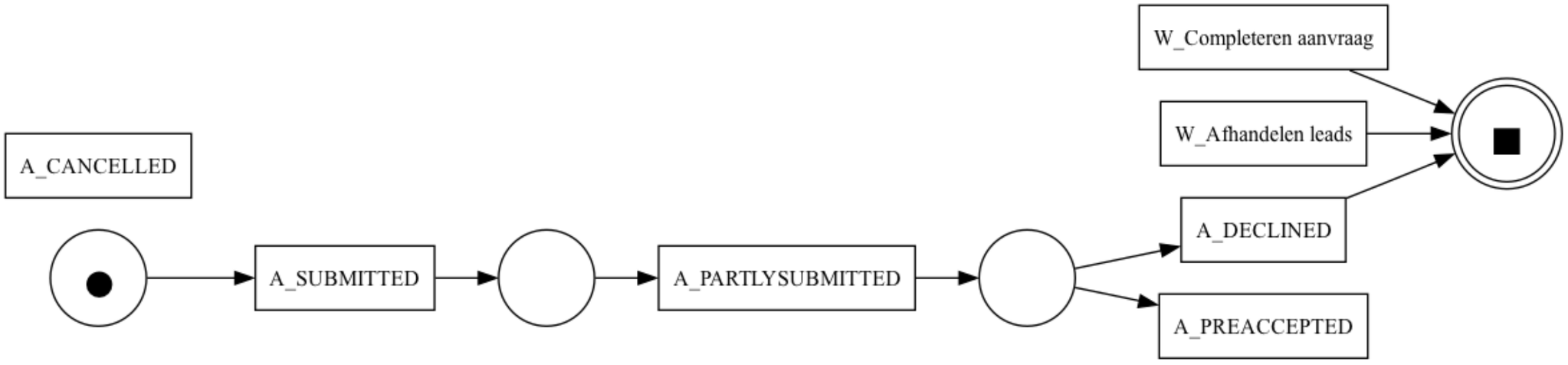
مجدداً با استفاده از الگوریتم‌های مختلف، مدل فرآیندی بر اساس گزارش رویداد مفروض، کشف شده و سپس شاخص‌های انطباق آن مورد بررسی و تحلیل واقع می‌گردد.



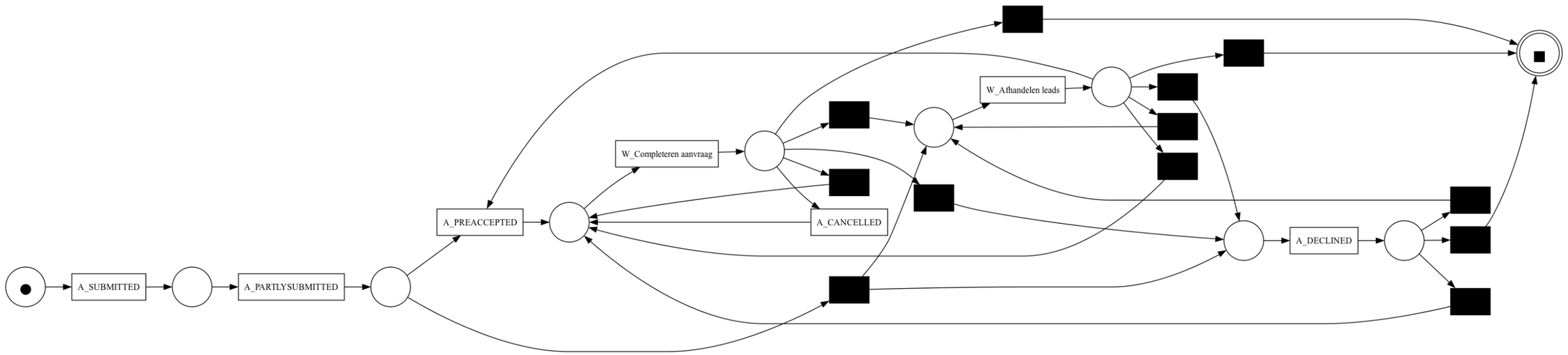
شکل ۱۶ - کشف مدل فرآیندی از طریق الگوریتم *Heuristics Miner*



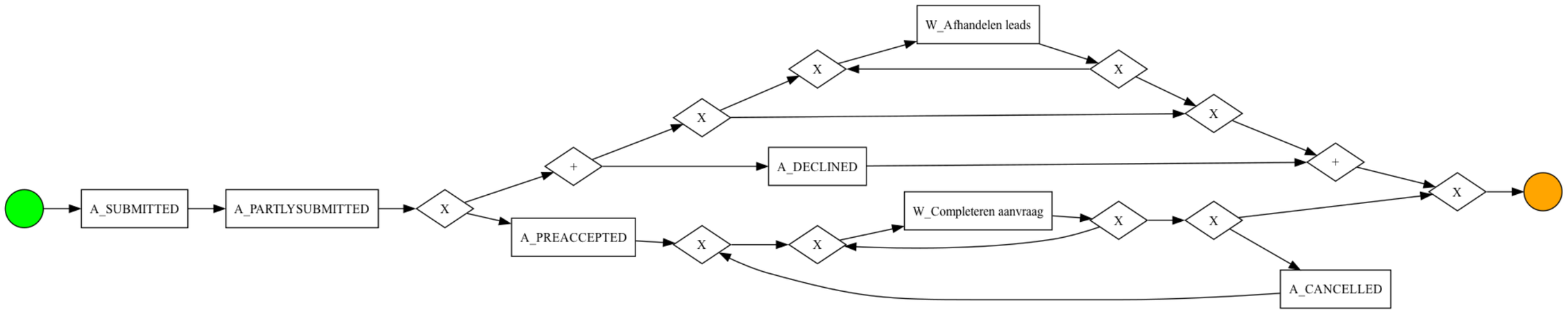
شکل ۱۷ - کشف مدل فرآیندی از طریق الگوریتم *Heuristics Miner – Petri Net*



شکل ۱۸ - کشف مدل فرآیندی از طریق *Alpha Algorithm — Petri Net*



شکل ۱۹ - کشف مدل فرآیندی از طریق *Petri Net Heuristics*



شکل ۲۰ - کشف مدل فرآیندی از طریق *Inductive Miner – BPMN*

۵-۲- بررسی انطباق فرآیند

در ادامه، کلیه‌ی شاخص‌ها نامبرده برای الگوریتم‌های به کار رفته در قالب جدول ۲۹ ارائه شده است:

جدول ۲۹- محاسبه‌ی شاخص‌های انطباق فرآیند کشف شده

Algorithm	Fitness	Percision	Generalization	Simplicity
Heuristics Miner	0.98	0.77	0.97	0.58
Alpha Algorithm	0.82	0.38	0.98	1.00
Petri Net Heuristics	-	0.92	-	0.57

الگوریتم آلفا اصلاً مطلوب به نظر نمی‌رسد. همچنین مجدداً میزان بالا بودن شاخص اول، نشان از بیش‌برازش[1] مدل است. از آن‌جایی که در همه‌ی مدل‌های کشف‌شده شاخص تعمیم‌پذیری بسیار بالاست، دقت مدل کاهش یافته است.

[1]. Over Fitness

فصل سوم: فرآیندکاوی

۱- دوباره‌کاری

۱-۱- تحلیل دوباره‌کاری درخواست‌ها

با توجه به ساختار کلی فرآیند و تکرر برخی فعالیت‌ها در هر درخواست به دفعات مختلف، بررسی این دوباره‌کاری‌ها امری ضروری جهت تحلیل فعالیت‌های متفاوت از منظر جنس فعالیت و همچنین منابع مورد نیاز آن است. از این رو در گام اول تلاش شد تا برای هر درخواست، فعالیت‌هایی که بیش از یک بار انجام شده مورد بررسی قرار گیرد تا بتوان در جدول اطلاعات درخواست‌ها برای هر پرونده لیستی از فعالیت‌های با بیش از یک تکرار ثبت کرد. علاوه بر این برای هر پرونده تعداد فعالیت‌های با بیشتر از یک تکرار نیز ثبت شد تا بتوان بر اساس این ویژگی رکوردها را مورد چینش مجدد قرار داد. در ادامه ۱۰ درخواست با بیشترین فعالیت‌های تکراری را در جدول ۳۰ می‌توان مشاهده نمود.

جدول ۳۰- محاسبه‌ی دوباره‌کاری فعالیت‌ها در هر درخواست

Case_ID	Amount_Request	Activities with Reworks	Number of Activities with Reworks
178843	10000	[k, l, m, n, p, r, s, t, u, v, x]	11
179885	35000	[k, l, m, n, p, r, s, t, u, v]	10
189280	7500	[k, l, m, n, p, r, s, t, u, v]	10
204859	32000	[k, l, m, n, p, r, s, t, u, v]	10
184087	7000	[k, l, m, n, p, r, s, t, u, v]	10
179899	3500	[k, l, m, n, r, s, t, u, v, w]	10
187076	7500	[k, l, m, n, p, r, s, t, u, v]	10
202740	3500	[k, l, m, n, p, r, s, t, u, v]	10
206135	10000	[k, l, m, n, p, r, s, t, u, v]	10
183471	25000	[k, l, m, n, p, r, s, t, u, v]	10
196623	23500	[k, l, m, n, p, r, s, t, u, v]	10
204442	6000	[k, l, m, n, p, r, s, t, u, v]	10
190956	15000	[k, l, m, n, p, r, s, t, u, v]	10
203206	15500	[k, l, m, n, p, r, s, t, u, v]	10
177206	21000	[k, l, m, p, r, s, t, u, v, w]	10
192115	30000	[k, l, m, n, p, r, s, t, u, v]	10
199165	30000	[k, l, m, n, p, r, s, t, u, v]	10
206937	13500	[k, l, m, n, p, r, s, t, u, v]	10
204742	15000	[k, l, m, n, p, r, s, t, u, v]	10
213738	10000	[k, l, m, n, p, r, s, t, u, v]	10

۱-۲- دلایل بروز دوباره‌کاری

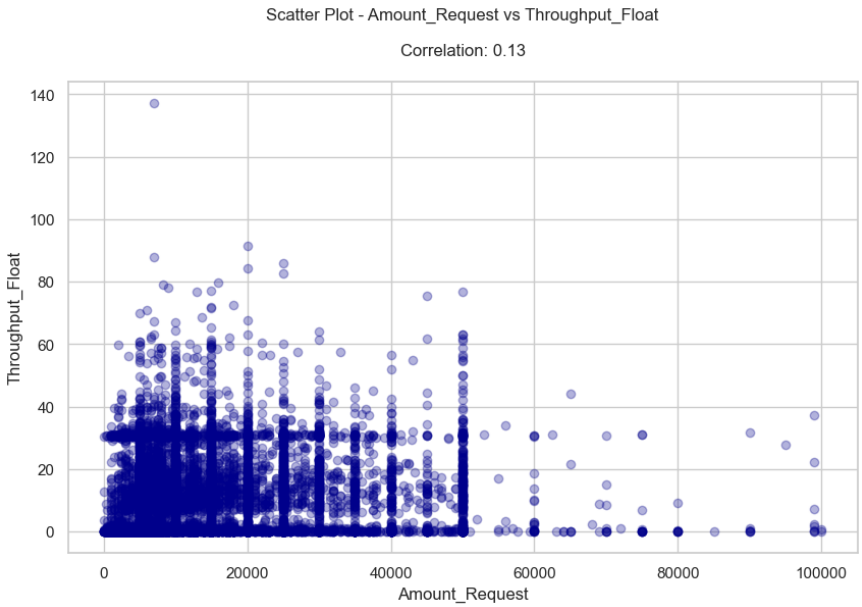
با توجه به جدول بخش قبل، روند خاصی را در بین مبلغ درخواستی پرونده‌های مختلف و میزان دوباره‌کاری آن‌ها نمی‌توان مشاهده کرد و اکثر مبالغ حول میانگین و با کمی اختلاف پراکنده شده است. اما آن‌چه جالب توجه است تکرر هر فعالیت به عنوان دوباره‌کاری در پرونده‌های مختلف می‌باشد. با توجه به جدولی که در ادامه ارائه شده است، به وضوح می‌توان مشاهده نمود که ۴ فعالیت $\{s, t, r, u\}$ با اختلاف بالا و به ترتیب ۷۳۶۷، ۵۰۱۱، ۴۷۵۵ و ۳۲۱۰ مرتبه به عنوان فعالیت شامل دوباره‌کاری در میان پرونده‌ها مشاهده شده است.

جدول ۳۱- فراوانی فعالیت‌هایی با بیشترین دوباره‌کاری

No.	Feature	Count
0	s	7367
1	t	5011
7	r	4755
2	u	3210
8	v	1647
3	k	1438
4	l	1438
5	m	1438
6	p	749
9	n	197
10	w	108
11	x	4

۲- تحلیل همبستگی

با توجه به آن که دو متغیر مبلغ وام درخواستی و همچنین مدت زمان اجرای فرآیند از نوع مقادیر عددی هستند؛ می‌توان برای بررسی رابطه‌ی آن‌ها از ضریب همبستگی پیرسون استفاده کرد که مقداری برابر با ۰.۱۳ دارد. این مقدار در محاسبات حاکی از عدم وابستگی این دو متغیر با یکدیگر است و می‌توان از این مقدار مثبت ناچیز چشم‌پوشی کرد. همچنین می‌توان با استفاده از نمودار نقطه‌ای ارائه شده در شکل ۲۱ به این موضوع صحه‌گذاری نمود.



شکل ۲۱ – نمودار نقطه‌ای مدت زمان اجرای هر درخواست بر حسب مبلغ وام

۳- خوشه‌بندی درخواست‌ها

همان‌طور که پیش‌تر محاسبه شد، میانگین مدت زمان اجرای فرآیند مقداری برابر ۸ روز و ۱۴ ساعت و ۵۵ دقیقه دارد و برای استخراج پرونده‌هایی با بیش از این میزان زمان کافی است تا روی جدول فرآیندها این شرط بررسی شود. بعد از بررسی شرط و استخراج پرونده‌های با مدت زمان اجرای بالا می‌توان جدولی جدید شامل مقادیر احتمالی موثر بر خوشه‌بندی پرونده‌ها طراحی نمود تا این عملیات را بتوان با بالاترین دقت روی مجموعه‌داده انجام داد.

جدول ۳۲ - مقادیر ورودی به مدل خوشه‌بندی به همراه برچسب مسیرها

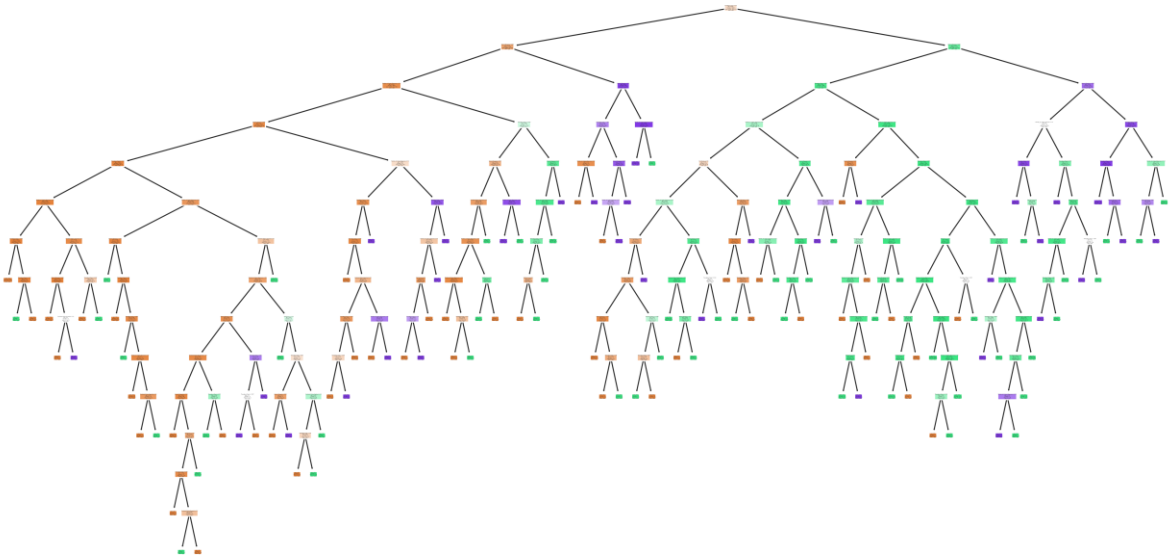
No	End Event	Amount Request	Throughput Float	Number of Resources Unique	Combined Status	Cluster
0	u	0.355763	-0.82253	5	Accepted	0
1	u	-0.89236	-1.11052	5	Accepted	0
2	x	-0.72594	10.54846	10	Accepted	1
7	i	-0.39311	0.868536	3	Cancelled	1
9	u	2.435965	-0.85623	5	Declined	2
…	…	…	…	…	…	…

حال پس از آماده‌سازی این جدول جهت اطمینان از عملکرد صحیح و سریع مدل خوشه‌بندی منتخب، مقادیر عددی جدول که دو متغیر مبلغ وام درخواستی و مدت زمان اجرای فرآیند است، استانداردسازی می‌شود. برای این امر با فرض نرمال بودن مقادیر این دو متغیر، از رابطه‌ی استانداردسازی منحنی نرمال استاندارد استفاده شده است. علاوه بر این پیش از خوشه‌بندی مجموعه‌ی داده‌ی استخراج‌شده، متغیرهای شامل مقادیر کیفی که فعالیت پایانی پرونده، تعداد منابع منحصر به فرد مورد استفاده در پرونده، وضعیت پایانی پرونده (پذیرش، رد، لغو و عدم تصمیم‌گیری) هستند، نیز وان‌هات انکد[1] شدند تا الگوریتم خوشه‌بندی بدون هیچ خللی اجرا شود. در گام پایانی یک مدل خوشه‌بندی کا-مین[2] با استفاده از سه خوشه و انجام ده اجرای مقدماتی پیاده‌سازی شد و در ادامه یک درخت تصمیم بر اساس این امر ترسیم شد که با توجه به خروجی مدلسازی آن می‌توان مشاهده نمود که موثرترین متغیرها در خوشه‌بندی این

[1]. One-hot Encode

[2]. K-Means

مجموعه‌ی داده به ترتیب مدت زمان فرآیند، مبلغ وام و تعداد منابع منحصر به فرد مورد استفاده است که منجر به طراحی این خوشه‌ها شده است.



شکل ۲۲ – درخت تصمیم مبتنی بر خوشه‌ی یادگیری شده

فصل چهارم: منابع و مراجع

[1] Moreira, C., Haven, E., Sozzo, S., & Wichert, A. (2018). Process mining with real world financial loan applications: Improving inference on incomplete event logs. PLoS One, 13(12), e0207806.

[2] Van der Aalst, W. M. (2013). Business process management: a comprehensive survey. International Scholarly Research Notices.

[3] Van Der Aalst, W. M. (2004). Business process management demystified: A tutorial on models, systems and standards for workflow management (pp. 1-65). Springer Berlin Heidelberg.

[4] Weske, M. (2007). Concepts, languages, architectures. Business Process Management.

[5] Van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. IEEE transactions on knowledge and data engineering, 16(9), 1128-1142.

[6] Van der Aalst, W. M. (2014). Process mining in the large: a tutorial. Business Intelligence: Third European Summer School, eBISS 2013, Dagstuhl Castle, Germany, July 7-12, 2013, Tutorial Lectures 3, 33-76.

[7] Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

[8] Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.

[9] Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53, 1487-1509.

[10] Ferreira, D. R. (2017). A primer on process mining: Practical skills with python and graphviz. Cham: Springer International Publishing.

[11] Van der Aalst, W. M., & Carmona, J. (2022). Process mining handbook (p. 503). Springer Nature.