FORE SCHOOL OF MANAGEMENT
NEW DELHI

Academic Session: 2023-2025

Machine Learning for Managers – II
"Employee Classification on the basis of Cluster Data by using Cross-validation and Ensemble Learning"

PGDM 32 Section: A

Submitted to:
Prof. Amarnath Mitra

Submitted by:
Arushi Khanna
Roll No.: 321012

TABLE OF CONTENTS

1. Project Objective

→ The first objective is segmentation of Employee Compensation Data using Cross-Validation.
→ The second objective is segmentation of Employee Compensation Data using Ensemble Learning
→ The third objective is to determine the appropriate classification model.
→ The fourth objective is to identify significant variables or features and their thresholds for classification.

2. Data Description

2.1 Dimension of Data
2.1.1 Data Source: https://www.kaggle.com/datasets/san-francisco/sf-employee-compensation
2.1.2 Data Size: 26 MB (Kaggle), 153 MB (Excel csv File)
2.1.3 Number of Variables: The number of variables in the csv file is 22.
2.1.4 Number of records: The number of records in the csv file is 6,83,277 (excluding naming column).

2.2 Description of Variables
Index Variables:
Organization Group Code: Gives the organisation groups an identification
Job Family Code: Gives the Job families an identification
Job Code: Gives the particular from a Job Family an identification
Year: The particular year for the variables and records
Department Code: An identifier for a department in an organisation
Union Code: An identifier for a particular union
Employee Identifier: An employee's identification code

Categorical Variables:

Year Type
Organization Group
Department
Union
Job Family
Job

Non – Categorical Variables:

Salaries
Overtime
Other Salaries
Total Salary
Retirement
Health and Dental
Other Benefits
Total Benefits
Total Compensation

2.3 **Descriptive Statistics**

2.3.1. Descriptive Statistics of Outcome Categorical Variables

It provides the statistics of cluster variable (categorical variable) by giving frequency as well as relative frequency

S Cluster	I Count (Cluster)	D Relative Frequency (Cluster)
cluster_1	38616	0.471
cluster_2	35895	0.438
cluster_0	7482	0.091

2.3.2. Descriptive Statistics of Input Categorical Variables

2.3.2.1 Year type

S Year Type	I Count (Year Type)	D Relative Frequency (Year Type)
Fiscal	46129	0.563
Calendar	35864	0.437

2.3.2.2 Organisation Group

S Organization Group	I Count (Organization Group)	D Relative Frequency (Organization Group)
Public Works, Transportation & Commerce	25971	0.317
Community Health	17775	0.217
Public Protection	15890	0.194
Culture & Recreation	7553	0.092
General Administration & Finance	7444	0.091
Human Welfare & Neighborhood Develo...	7301	0.089
General City Responsibilities	59	0.001

2.3.2.3 Department

S Department	I Count (Department)	D Relative Frequency (Department)
Sheriff	1003	0.012
Administrative Services	974	0.012
WTR Water Enterprise	874	0.011
Public Library	826	0.01
WWE Wastewater Enterprise	638	0.008
CRT Superior Court	620	0.008
HHP Hetch Hetchy Water & Po...	437	0.005
Trial Courts	428	0.005
JUV Juvenile Probation	372	0.005
DBI Building Inspection	357	0.004
CON Controller	348	0.004
DT GSA - Technology	335	0.004
HRD Human Resources	334	0.004
CAT City Attorney	333	0.004
PRT Port	332	0.004
DAT District Attorney	330	0.004
Registrar	330	0.004
DEM Emergency Management	312	0.004
CPC City Planning	301	0.004
REG Elections	299	0.004
FAM Fine Arts Museum	296	0.004
TTX Treasurer/Tax Collector	286	0.003
Controller	275	0.003
District Attorney	272	0.003
City Attorney	264	0.003
Department of Technology	263	0.003
Juvenile Court	255	0.003
Port	246	0.003
Human Resources	241	0.003
Dept of Emergency Management	239	0.003
City Planning	238	0.003
Building Inspection	233	0.003
Public Defender	196	0.002
ASR Assessor / Recorder	194	0.002
ADP Adult Probation	186	0.002
Treasurer/Tax Collector	184	0.002
PDR Public Defender	177	0.002

S Department	I Count (Department)	D Relative Frequency (Department)
PDR Public Defender	177	0.002
Fine Arts Museum	175	0.002
Adult Probation	173	0.002
Assessor	157	0.002
ENV Environment	151	0.002
BOS Board Of Supervisors	146	0.002
Mayor	146	0.002
MYR Mayor	137	0.002
ECN Economic & Wrkfrce Dvlpmnt	121	0.001
RET Retirement System	116	0.001
WAR War Memorial	114	0.001
Environment	114	0.001
CSS Child Support Services	110	0.001
War Memorial	110	0.001
Economic Workforce Developm...	105	0.001
Homeless Services	103	0.001
HSS Health Service System	99	0.001
Board Of Supervisors	96	0.001
Retirement Services	83	0.001
AAM Asian Art Museum	82	0.001
Child Support Services	72	0.001
ART Arts Commission	67	0.001
CHF Children;Youth & Families	66	0.001
Children Youth & Families	64	0.001
RNT Rent Arbitration Board	63	0.001
Asian Art Museum	63	0.001
GEN General City / Unallocated	59	0.001
Health Service System	53	0.001
Art Commission	46	0.001
Dept of Police Accountablility	45	0.001
Emergency Communications Dept	43	0.001
Rent Arbitration Board	37	0
HOM HOMELESSNESS SERVICES	36	0
HRC Human Rights Commission	30	0
CFC Children & Families Commsn	24	0
Human Rights Commission	23	0
Ethics Commission	21	0

S Department	I Count (Department)	D Relative Frequency (Department)
Ethics Commission	21	0
ETH Ethics Commission	20	0
SCI Academy Of Sciences	17	0
Children & Families Commission	16	0
WOM Status Of Women	15	0
Civil Service Commission	15	0
BOA Board Of Appeals - PAB	14	0
Dept Status of Women	12	0
Academy Of Sciences	10	0
CSC Civil Service Commission	8	0
Board Of Appeals	7	0
HHP CleanPowerSF	4	0
CII Commty Invest & Infrstrctr	2	0
LLB Law Library	1	0
Board of Appeals	1	0
Law Library	1	0

2.3.2.4 Union

S Union	I Count (Union)	D Relative Frequency (Union)
Court Unrepresente...	15	0
Members of the Boar...	15	0
Member, Board Of S...	14	0
Municipal Executive ...	13	0
Misc. Unrepresented...	13	0
Court-Unrep Profess...	12	0
Municipal Exec Assoc...	11	0
Elected Officials	8	0
Roofers and Waterp...	8	0
BrickLayers, Local 3	7	0
SEIU - Firefighter Pa...	6	0
Building Inspectors' ...	6	0
Court-Unrep Manag...	6	0
Roofers, Local 40	6	0
Court Unrepresente...	5	0
Hod Carriers, Local 166	5	0
Glaziers, Local 718	5	0
SEIU, Local 1021, H-1	5	0
Hod Carriers, Local 36	5	0
Institutional Police O...	4	0
Glaziers, Metal, and ...	4	0
Carpet, Linoleum an...	4	0
Carpet, Linoleum & S...	4	0
Bricklayers, Local 3	4	0
Laborers Int, Local 261	3	0
Executive Contract ...	2	0
Municipal Exec Assoc...	2	0
Law Librarian and As...	2	0
Municipal Executive ...	2	0
SF Courts Commissio...	1	0
Unrepresented Cont...	1	0
TWU Local 200	1	0
Port Director	1	0

2.3.2.5 Job Family

S Job Family	I Count (Job Family)	D Relative Frequency (Job Family)
Nursing	8288	0.101
Street Transit	6872	0.084
Police Services	5133	0.063
Journeyman Trade	4892	0.06
Human Services	4180	0.051
Public Service Aide	3747	0.046
Clerical, Secretarial & Steno	3576	0.044
Fire Services	3368	0.041
Housekeeping & Laundry	2672	0.033
Management	2500	0.03
Recreation	2475	0.03
Protection & Apprehension	2416	0.029
Professional Engineering	2395	0.029
Budget, Admn & Stats Analysis	2384	0.029
Correction & Detention	2284	0.028
Information Systems	2015	0.025
Legal & Court	1798	0.022
Med Therapy & Auxiliary	1682	0.021
Payroll, Billing & Accounting	1637	0.02
Library	1477	0.018
Lab, Pharmacy & Med Techs	1427	0.017
Personnel	1202	0.015
Sub-Professional Engineering	1120	0.014
Semi-Skilled & General Labor	1077	0.013
Agriculture & Horticulture	919	0.011
Airport Operation	865	0.011
Supervisory-Labor & Trade	804	0.01
Untitled	700	0.009
Skilled Labor	693	0.008
SF Superior Court	642	0.008
Medical & Dental	612	0.007
Community Development	606	0.007
Dietary & Food	591	0.007
Purchasing & Storekeeping	476	0.006
Construction Inspection	457	0.006
Probation & Parole	396	0.005
Pub Relations & Spec Assts	367	0.004

S Job Family	I Count (Job Family)	D Relative Frequency (Job Family)
Pub Relations & Spec Assts	367	0.004
Appraisal & Taxation	354	0.004
Energy & Environment	349	0.004
Public Health	339	0.004
Hospital Administration	323	0.004
Computer Operatns & Repro ...	308	0.004
Health & Sanitation Inspection	280	0.003
Public Safety Inspection	245	0.003
Park & Zoo	227	0.003
Revenue	203	0.002
Construction Project Mgmt	174	0.002
Museum & Cultural Affairs	132	0.002
Administrative-DPW/PUC	82	0.001
Administrative-Labor & Trades	50	0.001
MTA Operations	42	0.001
Property Administration	27	0
Administrative Secretarial	20	0
Port Operation	20	0
Emergency Coordination	20	0
Emergency Services	16	0
Administrative & Mgmt (Unrep)	15	0
Unassigned	12	0
SF Redevelopment Agency	10	0

2.3.2.6 Job

S Job	I Count (Job)	D Relative Frequency (Job)
Pr Administrative Analyst	395	0.005
Assoc Engineer	384	0.005
Librarian 1	382	0.005
Lieutenant, Fire Suppression	380	0.005
Senior Clerk Typist	363	0.004
Truck Driver	362	0.004
Food Service Worker	360	0.004
Administrative Analyst	354	0.004
Health Worker 2	351	0.004
Engineer	349	0.004
Automotive Mechanic	342	0.004
Manager II	331	0.004
Manager III	328	0.004
Museum Guard	324	0.004
Health Worker 3	322	0.004
Manager I	312	0.004
Senior Account Clerk	312	0.004
Electronic Maintenance Tech	307	0.004
Eligibility Worker	302	0.004
Deputy Court Clerk II	296	0.004
HSA Social Worker	285	0.003
Pool Lifeguard	284	0.003
Stationary Eng, Sewage Plant	275	0.003
Nursing Assistant	263	0.003
Senior Physician Specialist	255	0.003
Painter	242	0.003
PS Aide To Prof	242	0.003
Automotive Service Worker	241	0.003
IS Business Analyst-Principal	240	0.003
Stdntdsgntrain1, Arch/Eng/Plng	238	0.003
Accountant III	237	0.003
IS Business Analyst-Senior	237	0.003
Camp Assistant	232	0.003
Junior Administrative Analyst	230	0.003
Management Assistant	228	0.003
Lieutenant 3	228	0.003
Manager IV	226	0.003

S Job	I Count (Job)	D Relative Frequency (Job)
Manager IV	226	0.003
IS Engineer-Senior	222	0.003
Publ Svc Aide-Asst To Prof	219	0.003
Electrician	219	0.003
Principal Clerk	210	0.003
Transit Car Cleaner	210	0.003
Medical Social Worker	207	0.003
Physician Specialist	205	0.003
Deputy Probation Officer	204	0.002
Eng/Arch/Landscape Arch Sr	201	0.002
Asst Engr	200	0.002
Deputy Court Clerk III	199	0.002
Secretary 2	193	0.002
Manager V	189	0.002
Sheriff's Cadet	187	0.002
Junior Engineer	187	0.002
Deputy Sheriff (SFERS)	186	0.002
Program Specialist	184	0.002
Library Assistant	182	0.002
IT Operations Support Admn III	177	0.002
Public Safetycomm Disp	176	0.002
Automotive Machinist	174	0.002
Library Technical Assistant 1	174	0.002
Psychiatric Social Worker	173	0.002
Public SafetyComm Disp	173	0.002
Clerk Typist	162	0.002
Nurse Manager	162	0.002
Behavioral Health Clinician	162	0.002
Senior Personnel Analyst	160	0.002
StdntDsgnTrain1, Arch/Eng/Plng	160	0.002
Account Clerk	159	0.002
IS Engineer-Principal	159	0.002
Assistant Engineer	158	0.002
Pharmacy Technician	157	0.002
Carpenter	155	0.002
Eligibility Worker Supervisor	154	0.002
Construction Inspector	153	0.002

And so on and so forth

2.3.3 Descriptive Statistics: Non-Categorical Variables

2.3.3.1 Measures of Central Tendency

Row ID	Column	Min	Max	Mean	Std. d...	Variance	Skew...	Kurt...	Overall ...	No. miss...	No. NaNs	No. +oos	No. -oos	Median	Row count	Histogram
Salaries	Salaries	-17,635.32	584,297.61	69,578.331	47,721.087	2,277,302,190....	0.573	0.713	5,704,936,075...	0	0	0	0	?	81993	-17.635 584.298
Overtime	Overtime	-781.59	285,996.49	5,570.094	13,303.987	176,996,061.739	4.381	29.824	456,708,752.45	0	0	0	0	?	81993	-782 285.996
Other Salaries	Other Salar...	-554.6	548,318.35	3,829.925	8,726.299	76,148,294.323	13.524	523.559	314,027,007.24	0	0	0	0	?	81993	-555 548.318
Total Salary	Total Salary	-16,853.04	584,297.61	78,780.75	55,797.432	3,113,353,458....	0.644	0.469	6,459,469,996...	0	0	0	0	?	81993	-16.853 584.298
Retirement	Retirement	-15,350.26	114,934.16	13,699.821	10,103.507	102,080,858.831	0.524	0.679	1,123,289,439...	0	0	0	0	?	81993	-15.350 114.934
Health and Dental	Health and ...	-2,940.47	59,405.23	9,967.086	5,725.516	32,781,528.751	-0.604	-0.812	817,231,262.3	0	0	0	0	?	81993	-2.940 59.405
Other Benefits	Other Bene...	-3,513.86	35,148.15	5,037.76	3,770.684	14,218,055.027	1.01	3.007	413,061,027.22	0	0	0	0	?	81993	-3.514 35.148

Other Benefits	Other Bene...	-3,513.86	35,148.15	5,037.76	3,770.684	14,218,055.027	1.01	3.007	413,061,027.22	0	0	0	0	?	81993	-3.514 35.148
Total Benefits	Total Benefits	-7,509.22	151,645.18	28,704.667	17,832.791	318,008,437.164	-0.099	-0.601	2,353,581,729...	0	0	0	0	?	81993	-7.509 151.645
Total Compensa...	Total Comp...	-24,362.26	735,942.79	107,485.416	72,299.824	5,227,264,503....	0.408	0.025	8,813,051,726...	0	0	0	0	?	81993	-24.362 735.943

2.3.3.2 Correlation Statistics (using Test of Correlation)

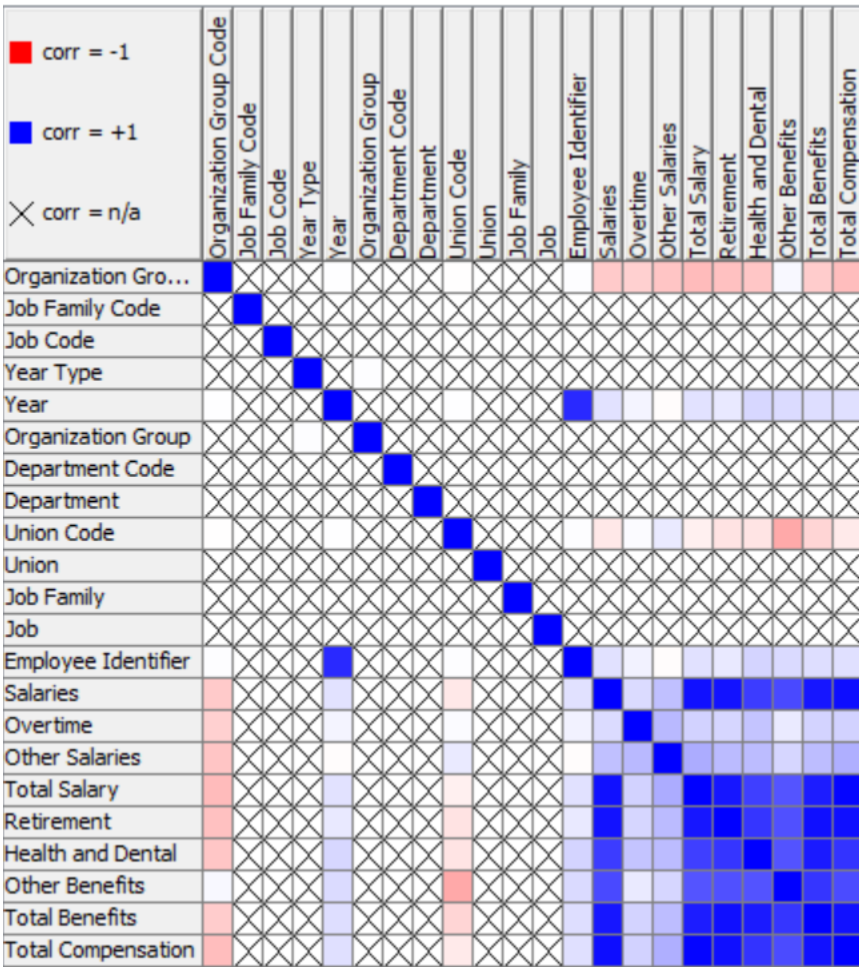
Row ID	First column name	Second...	Correlation value	p value	Degree...
Row0	Organization Group Code	Year	0.0038711295030974...	0.267662027116...	81991
Row1	Organization Group Code	Union Code	-0.0046714284381982...	0.181019253151...	81991
Row2	Organization Group Code	Employee Id...	0.005959528036041378	0.087921460275...	81991
Row3	Organization Group Code	Salaries	-0.2084251250690682	0.0	81991
Row4	Organization Group Code	Overtime	-0.1836090244690008	0.0	81991
Row5	Organization Group Code	Other Salaries	-0.22906720694098173	0.0	81991
Row6	Organization Group Code	Total Salary	-0.26536371144356125	0.0	81991
Row7	Organization Group Code	Retirement	-0.24124476021675945	0.0	81991
Row8	Organization Group Code	Health and ...	-0.22504839004079247	0.0	81991
Row9	Organization Group Code	Other Benefits	0.027312209514520586	5.107025913275...	81991
Row10	Organization Group Code	Total Benefits	-0.20074482027601478	0.0	81991
Row11	Organization Group Code	Total Compe...	-0.25738947147069474	0.0	81991
Row12	Job Family Code	Job Code	?	?	0
Row13	Job Family Code	Year Type	?	?	0
Row14	Job Family Code	Organizatio...	?	?	0
Row15	Job Family Code	Department ...	?	?	0
Row16	Job Family Code	Department	?	?	0
Row17	Job Family Code	Union	?	?	0
Row18	Job Family Code	Job Family	?	?	0
Row19	Job Family Code	Job	?	?	0
Row20	Job Code	Year Type	?	?	0
Row21	Job Code	Organizatio...	?	?	0
Row22	Job Code	Department ...	?	?	0
Row23	Job Code	Department	?	?	0
Row24	Job Code	Union	?	?	0
Row25	Job Code	Job Family	?	?	0
Row26	Job Code	Job	?	?	0
Row27	Year Type	Organizatio...	0.006204492327791245	0.788979108505...	6
Row28	Year Type	Department ...	?	?	0
Row29	Year Type	Department	?	?	0
Row30	Year Type	Union	?	?	0
Row31	Year Type	Job Family	?	?	0
Row32	Year Type	Job	?	?	0
Row33	Year	Union Code	0.005116740647162085	0.142883747589...	81991
Row34	Year	Employee Id...	0.8390548350603481	0.0	81991
Row35	Year	Salaries	0.11408972648890578	0.0	81991
Row36	Year	Overtime	0.04404004591696475	0.0	81991

Row ID	First column name	Second...	Correlation value	p value	Degree...
Row37	Year	Other Salaries	-0.009953004092339249	0.004371756390...	81991
Row38	Year	Total Salary	0.11262159116212003	0.0	81991
Row39	Year	Retirement	0.08493760618137396	0.0	81991
Row40	Year	Health and ...	0.15876945215772556	0.0	81991
Row41	Year	Other Benefits	0.14023544745338992	0.0	81991
Row42	Year	Total Benefits	0.12763468316328627	0.0	81991
Row43	Year	Total Compe...	0.12020177877614915	0.0	81991
Row44	Organization Group	Department ...	?	?	0
Row45	Organization Group	Department	?	?	0
Row46	Organization Group	Union	?	?	0
Row47	Organization Group	Job Family	?	?	0
Row48	Organization Group	Job	?	?	0
Row49	Department Code	Department	?	?	0
Row50	Department Code	Union	?	?	0
Row51	Department Code	Job Family	?	?	0
Row52	Department Code	Job	?	?	0
Row53	Department	Union	?	?	0
Row54	Department	Job Family	?	?	0
Row55	Department	Job	?	?	0
Row56	Union Code	Employee Id...	0.00771606191782626	0.027143442108...	81991
Row57	Union Code	Salaries	-0.09142517523953043	1.099209862001...	81991
Row58	Union Code	Overtime	0.016904803607550787	1.292754887805...	81991
Row59	Union Code	Other Salaries	0.08367270814457999	0.0	81991
Row60	Union Code	Total Salary	-0.05845209459323222	5.522949619678...	81991
Row61	Union Code	Retirement	-0.10995142262341598	7.115038496577...	81991
Row62	Union Code	Health and ...	-0.10772118644592514	4.054062639206...	81991
Row63	Union Code	Other Benefits	-0.3388525309429467	0.0	81991
Row64	Union Code	Total Benefits	-0.1652263596430773	0.0	81991
Row65	Union Code	Total Compe...	-0.0807633776079481	1.059972396069...	81991
Row66	Union	Job Family	?	?	0
Row67	Union	Job	?	?	0
Row68	Job Family	Job	?	?	0
Row69	Employee Identifier	Salaries	0.11819665710605165	0.0	81991
Row70	Employee Identifier	Overtime	0.05076760332195987	0.0	81991
Row71	Employee Identifier	Other Salaries	-0.011778082378191132	7.444822144974...	81991
Row72	Employee Identifier	Total Salary	0.11585401717043209	0.0	81991
Row73	Employee Identifier	Retirement	0.08777966291114811	0.0	81991

Row ID	First column name	Second...	Correlation value	p value	Degree...
Row73	Employee Identifier	Retirement	0.08777966291114811	0.0	81991
Row74	Employee Identifier	Health and ...	0.1700565260747127	0.0	81991
Row75	Employee Identifier	Other Benefits	0.14685428352630214	0.0	81991
Row76	Employee Identifier	Total Benefits	0.13018778935306685	0.0	81991
Row77	Employee Identifier	Total Compe...	0.12260693772677879	0.0	81991
Row78	Salaries	Overtime	0.14120023587398486	0.0	81991
Row79	Salaries	Other Salaries	0.24606035831508835	0.0	81991
Row80	Salaries	Total Salary	0.9392123563079754	0.0	81991
Row81	Salaries	Retirement	0.9309625293967527	0.0	81991
Row82	Salaries	Health and ...	0.7659238528004219	0.0	81991
Row83	Salaries	Other Benefits	0.715150196253188	0.0	81991
Row84	Salaries	Total Benefits	0.9120386625710687	0.0	81991
Row85	Salaries	Total Compe...	0.9504234958763824	0.0	81991
Row86	Overtime	Other Salaries	0.2764025272384125	0.0	81991
Row87	Overtime	Total Salary	0.1780799401553044	0.0	81991
Row88	Overtime	Retirement	0.16097575722393156	0.0	81991
Row89	Overtime	Health and ...	0.23217316225151768	0.0	81991
Row90	Overtime	Other Benefits	0.0821365982459494	0.0	81991
Row91	Overtime	Total Benefits	0.17153656711761747	0.0	81991
Row92	Overtime	Total Compe...	0.17827561578145953	0.0	81991
Row93	Other Salaries	Total Salary	0.327114872596103	0.0	81991
Row94	Other Salaries	Retirement	0.2682020547655558	0.0	81991
Row95	Other Salaries	Health and ...	0.26095949016698833	0.0	81991
Row96	Other Salaries	Other Benefits	0.1614733419335772	0.0	81991
Row97	Other Salaries	Total Benefits	0.25742694847794195	0.0	81991
Row98	Other Salaries	Total Compe...	0.31537428063558437	0.0	81991
Row99	Total Salary	Retirement	0.913861032684394	0.0	81991
Row100	Total Salary	Health and ...	0.75163325125133	0.0	81991
Row101	Total Salary	Other Benefits	0.6715267312120912	0.0	81991
Row102	Total Salary	Total Benefits	0.8907250396453937	0.0	81991
Row103	Total Salary	Total Compe...	0.9803466083512917	0.0	81991
Row104	Retirement	Health and ...	0.7927011937453792	0.0	81991
Row105	Retirement	Other Benefits	0.6858720955214168	0.0	81991
Row106	Retirement	Total Benefits	0.9427367800106262	0.0	81991
Row107	Retirement	Total Compe...	0.9400321999114722	0.0	81991
Row108	Health and Dental	Other Benefits	0.6758154787795075	0.0	81991
Row109	Health and Dental	Total Benefits	0.8970208021446427	0.0	81991
Row109	Health and Dental	Total Benefits	0.8970208021446427	0.0	81991
Row110	Health and Dental	Total Compe...	0.7998229813946881	0.0	81991
Row111	Other Benefits	Total Benefits	0.7925736129275652	0.0	81991
Row112	Other Benefits	Total Compe...	0.7076584319118923	0.0	81991
Row113	Total Benefits	Total Compe...	0.9327580707805774	0.0	81991

Row ID	Organiz...	Job Fa...	Job Code	Year Type	Year	Organiz...	Depart...	Depart...	Union Code	Union	Job Fa...	Job	Employee Identifier
Organization ...	1.0	?	?	?	0.0038711295030974...	?	?	?	-0.0046714284381982...	?	?	?	0.005959528036041...
Job Family Code	?	1.0	?	?	?	?	?	?	?	?	?	?	?
Job Code	?	?	1.0	?	?	?	?	?	?	?	?	?	?
Year Type	?	?	?	1.0	?	0.00620449...	?	?	?	?	?	?	?
Year	0.00387112...	?	?	?	1.0	?	?	?	0.005116740647162085	?	?	?	0.8390548350603481
Organization ...	?	?	?	0.00620449...	?	1.0	?	?	?	?	?	?	?
Department C...	?	?	?	?	?	?	1.0	?	?	?	?	?	?
Department	?	?	?	?	?	?	?	1.0	?	?	?	?	?
Union Code	-0.0046714...	?	?	?	0.005116740647162085	?	?	?	1.0	?	?	?	0.007716061917826
Union	?	?	?	?	?	?	?	?	?	1.0	?	?	?
Job Family	?	?	?	?	?	?	?	?	?	?	1.0	?	?
Job	?	?	?	?	?	?	?	?	?	?	?	1.0	?
Employee Ide...	0.005959528...	?	?	?	0.8390548350603481	?	?	?	0.007716061917826	?	?	?	1.0
Salaries	-0.2084251...	?	?	?	0.11408972648890578	?	?	?	-0.09142517523953043	?	?	?	0.11819665710605165
Overtime	-0.1836090...	?	?	?	0.04404004591696475	?	?	?	0.016904803607550787	?	?	?	0.05076760332195987
Other Salaries	-0.2290672...	?	?	?	-0.009953004092339249	?	?	?	0.08367270814457999	?	?	?	-0.011778082378191...
Total Salary	-0.2653637...	?	?	?	0.11262159116212003	?	?	?	-0.05845209459323222	?	?	?	0.11585401717043209
Retirement	-0.2412447...	?	?	?	0.08493760618137396	?	?	?	-0.10995142262341598	?	?	?	0.08777966291114811
Health and D...	-0.2250483...	?	?	?	0.15876945215772556	?	?	?	-0.10772118644592514	?	?	?	0.1700565260747127
Other Benefits	0.02731220...	?	?	?	0.14023544745338992	?	?	?	-0.3388525309429467	?	?	?	0.14685428352630214
Total Benefits	-0.2007448...	?	?	?	0.12763468316328627	?	?	?	-0.1652263596430773	?	?	?	0.13018778935306685
Total Compen...	-0.2573894...	?	?	?	0.12020177877614915	?	?	?	-0.0807633776079481	?	?	?	0.12260693772677879

Row ID	Employee Identifier	Salaries	Overtime	Other Salaries	Total Salary	Retirement	Health and Dental	Other Benefits	Total Benefits	Total Compensa...
Organization ...	0595952803604 1...	-0.2084251250690...	-0.1836090244690...	-0.22906720694098...	-0.26536371144356...	-0.24124476021675...	-0.22504839004079...	0.027312209514520...	-0.20074482027601...	-0.25738947147069...
Job Family Code		?	?	?	?	?	?	?	?	?
Job Code		?	?	?	?	?	?	?	?	?
Year Type		?	?	?	?	?	?	?	?	?
Year	390548350603481	0.11408972648890...	0.04404004591696...	-0.00995300409233...	0.11262159116212003	0.08493760618137396	0.15876945215772556	0.14023544745338992	0.12763468316328627	0.12020177877614915
Organization ...		?	?	?	?	?	?	?	?	?
Department C...		?	?	?	?	?	?	?	?	?
Department		?	?	?	?	?	?	?	?	?
Union Code	07716061917826 26	-0.0914251752395...	0.01690480360755...	0.08367270814457999	-0.05845209459323...	-0.10995142262341...	-0.10772118644592...	-0.3388525309429467	-0.1652263596430773	-0.0807633776079481
Union		?	?	?	?	?	?	?	?	?
Job Family		?	?	?	?	?	?	?	?	?
Job		?	?	?	?	?	?	?	?	?
Employee Ide...		0.11819665710605...	0.05076760332195...	-0.01177808237819...	0.11585401717043209	0.08777966291114811	0.1700565260747127	0.14685428352630214	0.13018778935306685	0.12260693772677879
Salaries	1819665710605165	1.0	0.14120023587398...	0.24606035831508835	0.9392123563079754	0.9309625293967527	0.7659238528004219	0.715150196253188	0.9120386625710687	0.9504234958763824
Overtime	5076760332195987	0.14120023587398...	1.0	0.276402527238412 5	0.17807994015530 44	0.16097575722393 156	0.23217316225151768	0.082136598245949 4	0.17153656711761747	0.17827561578145953
Other Salaries	011778082378191...	0.24606035831508...	0.276402527238412 5	1.0	0.32711487259610 3	0.26820205476555 58	0.26095949016698833	0.16147334193357 72	0.25742694847794195	0.31537428063558437
Total Salary	1585401717043209	0.9392123563079754	0.17807994015530 44	0.32711487259610 3	1.0	0.91386103268439 4	0.75163325125133	0.671526731212091 2	0.8907250396453937	0.98034660835129 17
Retirement	8777966291114811	0.9309625293967527	0.16097575722393...	0.26820205476555 58	0.91386103268439 4	1.0	0.79270119374537 92	0.685872095521416 8	0.94273678001062 62	0.94003219991147 22
Health and D...	700565260747127	0.7659238528004219	0.23217316225151...	0.26095949016698833	0.75163325125133	0.7927011937453792	1.0	0.675815478779507 5	0.8970208021446427	0.7998229813946881
Other Benefits	4685428352630214	0.715150196253188	0.082136598245949 4	0.16147334193357 72	0.67152673121209 12	0.68587209552141 68	0.675815478779507 5	1.0	0.7925736129275652	0.7076584319118923
Total Benefits	3018778935306685	0.9120386625710687	0.17153656711761...	0.25742694847794195	0.8907250396453937	0.9427367800106262	0.8970208021446427	0.7925736129275652	1.0	0.9327580707805774
Total Compen...	2260693772677879	0.9504234958763824	0.17827561578145...	0.31537428063558437	0.9803466083512917	0.9400321999114722	0.7998229813946881	0.7076584319118923	0.9327580707805774	1.0



3. Analysis Of Data

3.1 Data Pre-Processing
3.1.1 Missing Data Statistics and Treatment
3.1.1.1 Missing Data Statistics: 0
3.1.1.2 Missing Data Treatment: 0
3.1.1.2.1 Removal of Records with More Than 50% Missing Data: None
3.1.1.3 Missing Data Statistics of categorical Variables: 0
3.1.1.3.1 Missing Data Treatment: Categorical Variables or Features: 0
3.1.1.3.1.1 Removal of Variables or Features with More Than 50% Missing Data: None
3.1.1.4 Missing Data Statistics of non-categorical Variables: 0
3.1.1.4.1 Missing Data Treatment of non-categorical Variables: 0
3.1.1.4.1.1 Removal of Variables or Features with More Than 50% Missing Data: None

3.1.2 Numerical Encoding of Variables

In this case, category to variable node will be used to encode categorical data into numbers such as:

Year Type:
Calendar = 0, Fiscal = 1

Organization Group:
Public Protection = 0
Public Works, Transportation & Commerce = 1
Human Welfare & Neighbourhood Development = 2
Community Health = 3
Culture & Recreation = 4
General Administration & Finance = 5

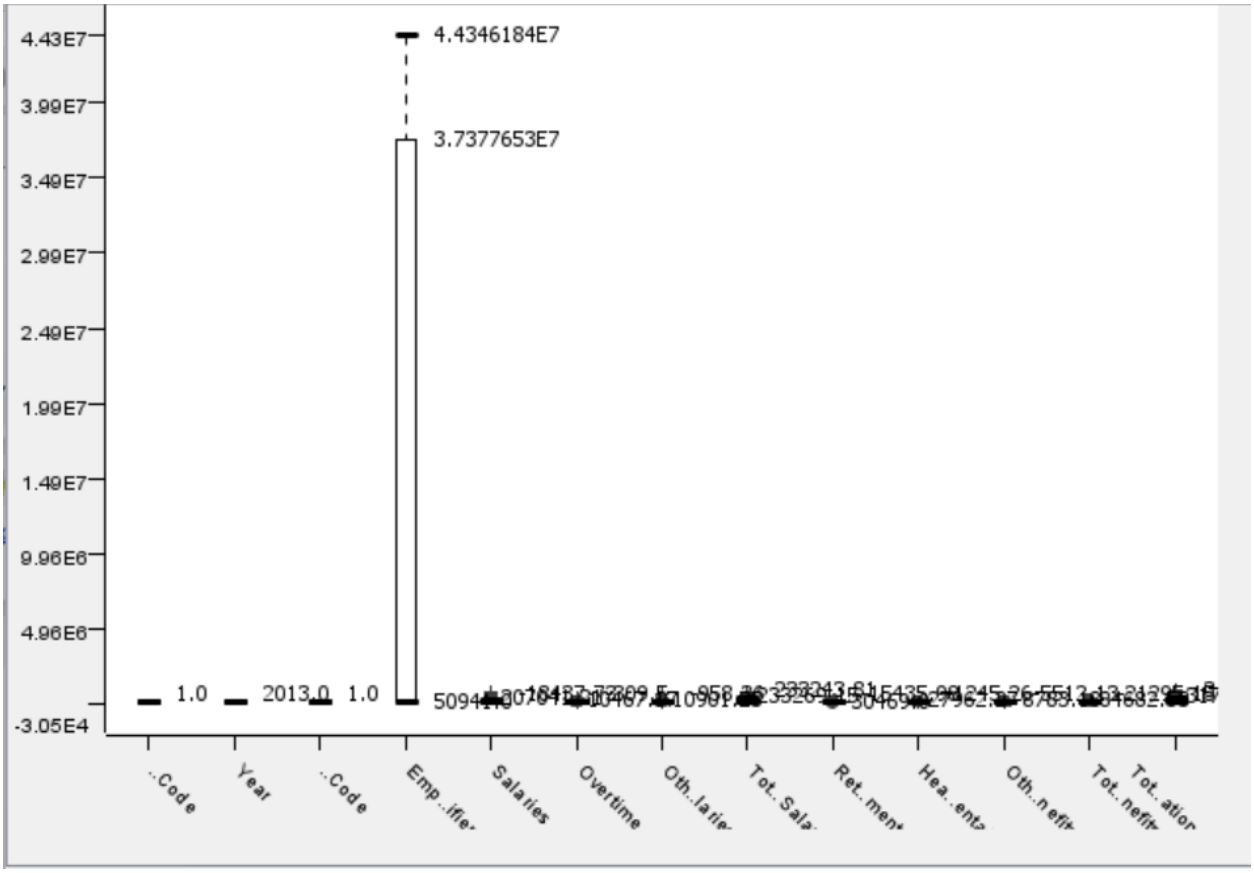
And so on and so forth

3.1.3 Outlier Statistics Treatment
3.1.3.1 Outlier Statistics: Non-Categorical Variables

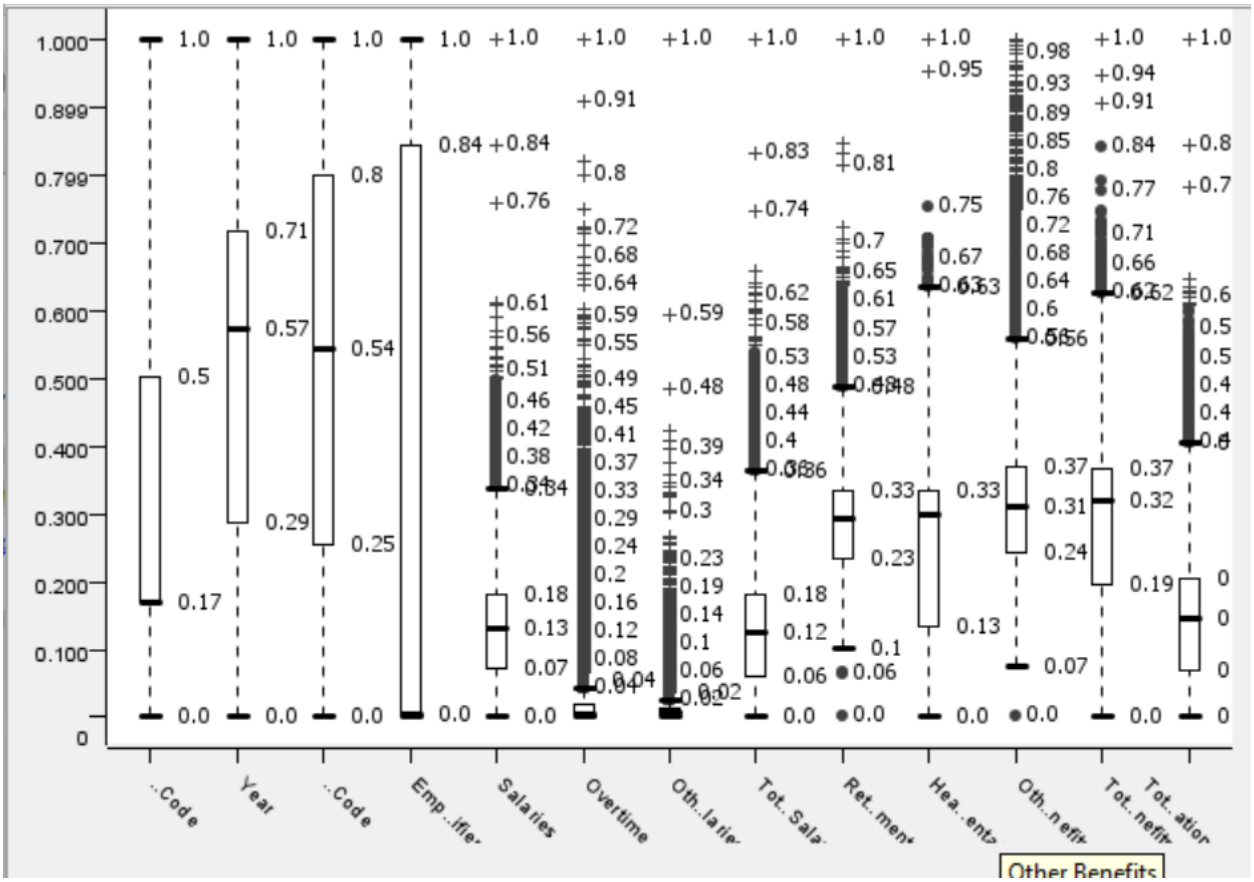
Row ID	S Outlier column	I Member count	I Outlier count	D Lower bound	D Upper bound
Row0	Salaries	81993	622	-0.1	0.374
Row1	Overtime	81993	12851	-0.02	0.04
Row2	Other Salaries	81993	8591	-0.011	0.021
Row3	Total Salary	81993	648	-0.121	0.418
Row4	Retirement	81993	589	-0.022	0.45
Row5	Health and Dental	81993	30	-0.096	0.494
Row6	Other Benefits	81993	866	-0.078	0.503
Row7	Total Benefits	81993	82	-0.151	0.578
Row8	Total Compensation	81993	350	-0.125	0.451

3.1.3.2 Normalization using Min-Max Scaler

Before Normalization

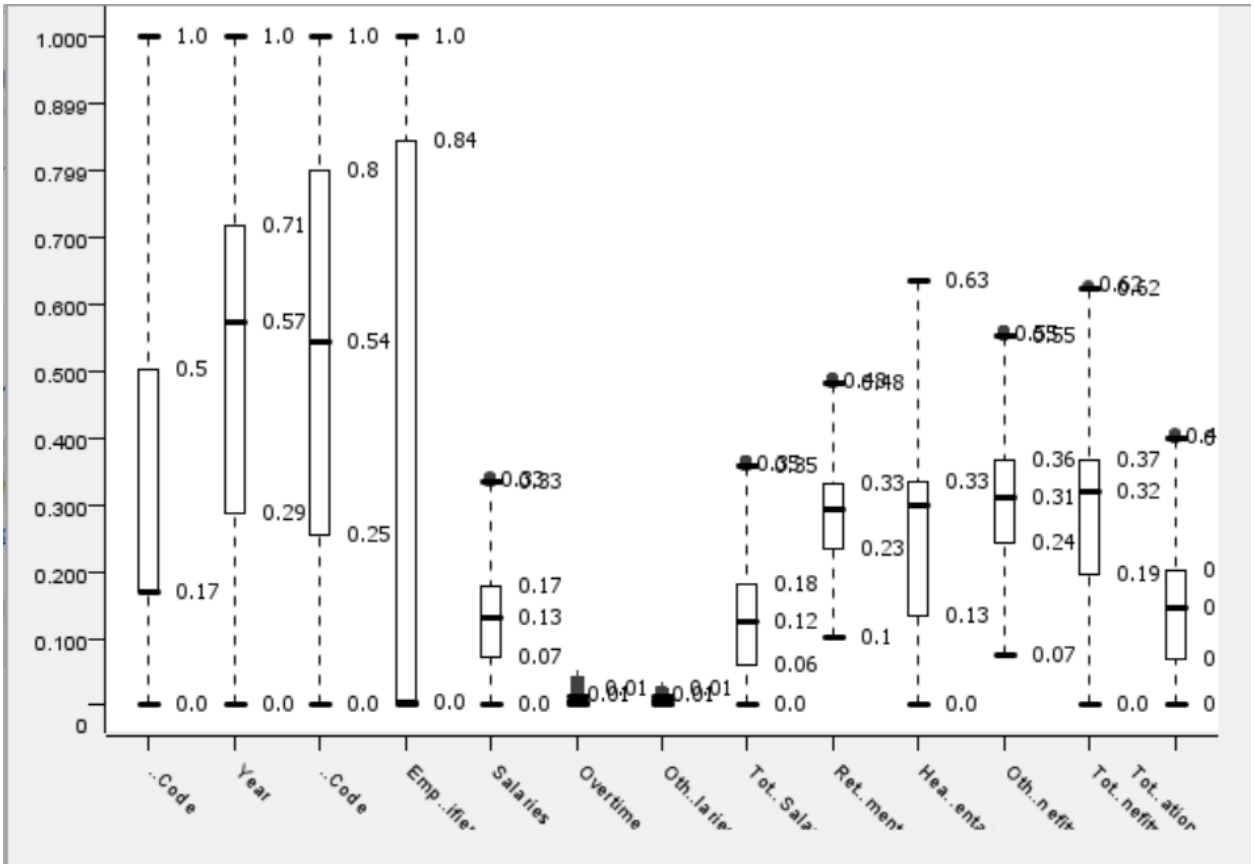


After Normalization

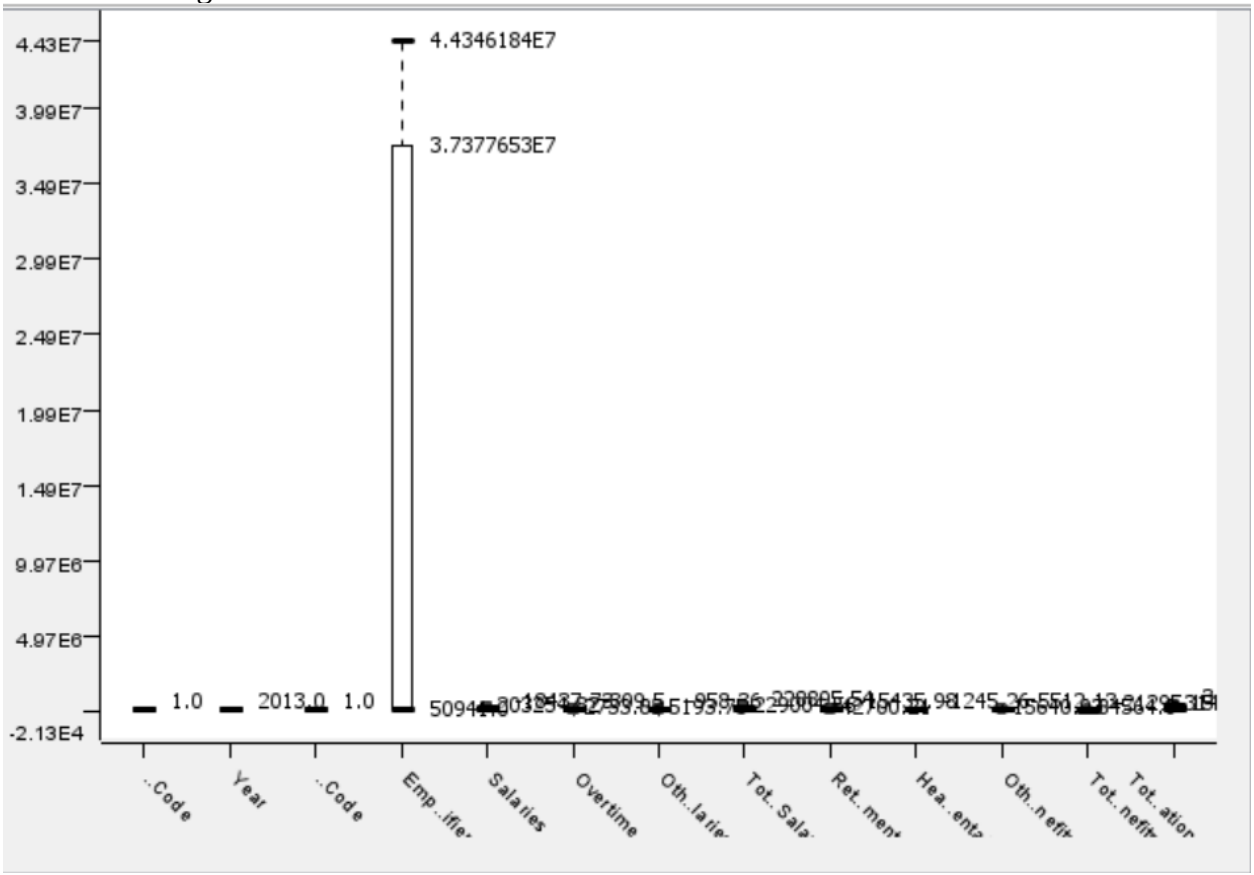

Year
Salaries
Overtime
Other Benefits

After treating Outliers

Outlier Treatment using replacement strategy where the values are replaced to the closest permitted value



De-Normalizing the Data



3.2. Data Analysis

3.2.1. Cross-Validation using Decision Tree

Cross-validation using a decision tree involves splitting the dataset into k subsets, training the decision tree on k-1 subsets and validating on the remaining subset by repeating this process k times and averaging the results to assess the model's performance and generalization ability.

3.2.2. Cross-Validation using Other Methods

3.2.2.1. Logistic Regression

Cross-validation with logistic regression involves partitioning the dataset into training and validation sets, fitting the logistic regression model on the training data and evaluating its performance on the validation set. This process is repeated multiple times with different partitions to estimate the model's generalization performance and minimize overfitting.

3.2.2.2. K-Nearest Neighbours

Cross-validation with KNN entails splitting the dataset into training and validation sets, then iterating through different values of k (number of nearest neighbours) to find the optimal k value that minimizes error on the validation set. This process helps assess the KNN model's performance and its ability to generalize to new data.

3.2.3. Ensemble Method using Random Forest

Random forest is an ensemble learning method where multiple decision trees are trained on random subsets of the data and features. During prediction, each tree votes on the outcome and the final prediction is determined by the majority vote. This approach improves prediction accuracy and reduces overfitting compared to individual decision trees.

3.2.4. Ensemble Method using XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that uses a gradient boosting framework. It sequentially builds multiple decision trees, each correcting the errors of the previous one. XGBoost incorporates regularization techniques to prevent overfitting and is known for its efficiency and effectiveness in various machine learning tasks.

3.2.1.1. Model Performance Evaluation of Cross-Validation using Decision Tree

Without Pruning

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	4608	0	5111	0	1	1	1	1	1	?	?
cluster_2	4167	0	5552	0	1	1	1	1	1	?	?
cluster_0	944	0	8775	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	4608	0	0
cluster_2	0	4167	0
cluster_0	0	0	944

With Pruning

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	4608	0	5111	0	1	1	1	1	1	?	?
cluster_2	4167	0	5552	0	1	1	1	1	1	?	?
cluster_0	944	0	8775	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	4608	0	0
cluster_2	0	4167	0
cluster_0	0	0	944

Cluster 1

• This cluster has a high number of true positives and true negatives indicating that the model correctly classified most instances within this cluster.
• The precision and recall scores are both very high suggesting that the model effectively identifies true positives while also minimizing false positives.

Cluster 2

• This cluster has a same recall and precision compared to cluster 1, indicating that the model's performance is as strong for this segment.
• There are 0 false positives and false negatives in both segments.
• Despite the strong performance metrics, the specificity is very high indicating that the model correctly identifies true negatives within this cluster.

Cluster 0

• This cluster has the same recall and precision.
• The number of false positives is relatively 0 suggesting that the model effectively minimizes misclassifications within this cluster.
• Specificity scores are high indicating that the model correctly identifies the true negatives within this cluster.

Comparative analysis of decision tree with and without pruning

• Pruning generally improves precision and specificity while slightly reducing recall and sensitivity.
• But as we can see here with or without pruning both the results are more or less same which means that for ensemble learning for both decision tree learners with or without pruning the results will be consistent and more or less optimised.

3.2.2.1.Model Performance Evaluation of Cross-Validation using Other Methods

Logistic Regression

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	2905	1683	20
cluster_2	2268	1894	5
cluster_0	625	317	2

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	2905	2893	2218	1703	0.63	0.501	0.63	0.434	0.558	?	?
cluster_2	1894	2000	3552	2273	0.455	0.486	0.455	0.64	0.47	?	?
cluster_0	2	25	8750	942	0.002	0.074	0.002	0.997	0.004	?	?
Overall	?	?	?	?	?	?	?	?	?	0.494	0.072

The confusion matrix shows the performance of a logistic regression model classifying data points into three classes: cluster_0, cluster_1 and cluster_2. Here's a breakdown of the table:

Values in each cell represent the count of data points that fall into that classification.

Looking at the values in the table:

- **2905:** Out of the data points that actually belong to cluster_1 (positive class), the model correctly classified 2905 of them (True Positives).
- **2893:** The model incorrectly classified 2893 data points that don't belong to cluster_1 (negative class) as cluster_1 (False Positives).
- **2218:** Out of the data points that don't belong to cluster_1, the model correctly classified 10 of them (True Negatives). This number is quite low compared to the other values, indicating the model might be struggling to identify points that are not cluster_1.

The same could be followed up for cluster_2 and cluster_0.

Overall Performance:

- Based on these values, the model seems to be good at identifying actual cluster_1 data points (high True Positives). However, it's also making a significant number of mistakes (False Positives) by classifying data points that don't belong to cluster_1 as cluster_1. It's also missing some actual cluster_1 data points (False Negatives).

K nearest Neighbor

K = 7

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	4539	176	4935	69	0.985	0.963	0.985	0.966	0.974	?	?
cluster_2	3984	269	5283	183	0.956	0.937	0.956	0.952	0.946	?	?
cluster_0	743	8	8767	201	0.787	0.989	0.787	0.999	0.877	?	?
Overall	?	?	?	?	?	?	?	?	?	0.953	0.919

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	4539	69	0
cluster_2	175	3984	8
cluster_0	1	200	743

K = 9

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	4550	199	4912	58	0.987	0.958	0.987	0.961	0.973	?	?
cluster_2	3964	279	5273	203	0.951	0.934	0.951	0.95	0.943	?	?
cluster_0	722	5	8770	222	0.765	0.993	0.765	0.999	0.864	?	?
Overall	?	?	?	?	?	?	?	?	?	0.95	0.913

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	4550	58	0
cluster_2	198	3964	5
cluster_0	1	221	722

K = 19

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	4547	227	4884	61	0.987	0.952	0.987	0.956	0.969	?	?
cluster_2	3936	351	5201	231	0.945	0.918	0.945	0.937	0.931	?	?
cluster_0	654	4	8771	290	0.693	0.994	0.693	1	0.816	?	?
Overall	?	?	?	?	?	?	?	?	?	0.94	0.895

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	4547	61	0
cluster_2	227	3936	4
cluster_0	0	290	654

In KNN, the number of neighbors to be considered are from K = 7, K = 9 and K = 19. From the images, it is seen that as the number of k increases the accuracy also increases. For K=7, as the accuracy is the highest from all the other k's, this cluster will be considered.

Cluster_1

- True Positives: 4539, False Positives: 269, True Negatives: 4935, False Negatives: 69
- Recall: 0.985, Precision: 0.937, Sensitivity: 0.985, Specificity: 0.966
- F-measure: 0.974, Accuracy: 0.953

In cluster_0, the KNN model achieved high recall indicating that it effectively identifies true positives within this cluster. However, the precision is relatively high emphasizing a lower rate of false positives. The model's specificity is extremely high indicating that it nicely identifies true negatives.
The overall accuracy is high which shows that the model's performance may vary across different metrics.

Cluster _2
• True Positives: 3984, False Positives: 269, True Negatives: 5283, False Negatives: 183
• Recall: 0.4956 , Precision: 0.983, Sensitivity: 0.956, Specificity: 0.952
• F-measure: 0.946, Accuracy: 0.953

Cluster _0
• True Positives: 743, False Positives: 8, True Negatives: 8767, False Negatives: 201
• Recall: 0.787, Precision: 0.989, Sensitivity: 0.787, Specificity: 0.999
• F-measure: 0.877, Accuracy: 0.953 In cluster_0 the KNN model has low recall but precision indicating that it struggles to correctly classify instances within this cluster. However, the model exhibits high specificity showing a strong ability to identify true negatives. The overall accuracy is moderate reflecting the model's mixed performance across different metrics. The overall accuracy of the KNN model is moderate showing mixed performance across different clusters. However, Cohen's Kappa coefficient suggests very high agreement beyond chance among the predicted and actual cluster labels

3.2.3.1. Model Performance Evaluation of Random Forest

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	851	253	769	71	0.923	0.771	0.923	0.752	0.84	?	?
cluster_2	683	114	997	150	0.82	0.857	0.82	0.897	0.838	?	?
cluster_0	28	15	1740	161	0.148	0.651	0.148	0.991	0.241	?	?
Overall	?	?	?	?	?	?	?	?	?	0.803	0.645

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	851	65	6
cluster_2	141	683	9
cluster_0	112	49	28

Cluster 1
• Cluster 1 exhibits extremely high-performance metrics, with almost perfect recall, precision, sensitivity and specificity.
• The model effectively identifies true positives while minimizing false positives and false negatives, indicating robust predictive power.
• Employees in this cluster are likely to have characteristics that make them highly reliable for work, resulting in minimal disadvantages.

Cluster 2
• Cluster 2 exhibits lower performance metrics compared to cluster 1, with good recall, precision, and F-measure.
• The model correctly identifies a significant portion of true positives but has a lower rate of false positives and false negatives.
• Employees in this cluster certainly portray less disadvantages than before.

Cluster
• Cluster 0 demonstrates low performance metrics.
• The model effectively identifies true positives while maintaining a low false positive rate, suggesting reliable predictions for employees.
• This cluster has performed well at identifying true negatives with a significantly higher rate of those than other.

3.2.3.2. Model Performance Evaluation of XGBoost

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	15377	0	1022	0	1	1	1	1	1	?	?
cluster_2	833	0	15566	0	1	1	1	1	1	?	?
cluster_0	189	0	16210	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	15377	0	0
cluster_2	0	833	0
cluster_0	0	0	189

1. Cluster 1
• This cluster has a high number of true positives (15377) and a zero false positive rate, indicating that the model is very good at correctly identifying positive cases while minimizing false alarms.
• The recall, precision, sensitivity and F-measure are all very high, indicating excellent performance in correctly identifying positive cases and minimizing false positives.
• The specificity is also high at 1 indicating a low false positive rate.

2. Cluster 2
• This cluster has a lower number of true positives (833) compared to the other, and a zero false positive rate, indicating that the model's performance in identifying positive cases is not as strong in this cluster.
• The recall, precision, sensitivity and F-measure are all high, indicating that

the model's performance in correctly identifying positive cases and minimizing false positives is same.
• The specificity is still high at 1 indicating a relatively low false positive rate.

3. Cluster 0
• This cluster has low number of true positives (189) and a zero low false positive rate indicating good performance in correctly identifying positive cases while minimizing false alarms.
• The recall, precision, sensitivity and F-measure are all high, suggesting that the model performs well in correctly identifying positive cases and minimizing false positives.
• The specificity is also high at 1 indicating a low false positive rate.

3.3. Variable or Feature Analysis for Decision Tree

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
cluster_1	4608	0	5111	0	1	1	1	1	1	?	?
cluster_2	4167	0	5552	0	1	1	1	1	1	?	?
cluster_0	944	0	8775	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	4608	0	0
cluster_2	0	4167	0
cluster_0	0	0	944

Overall Analysis

All the metrics are at the highest indicating the model is performing well with 0 False Positives and False Negatives this model is correctly identifying everything and in general all the variables used would be significant.

4. Results and Observations

4.1. Comparing Supervised Learning models: Cross Validation using Decision Tree VS Cross Validation using Logistic Regression, KNN

Cross validation using Decision Tree

No pruning

Cluster \P...	cluster_1	cluster_2	cluster_0
cluster_1	4608	0	0
cluster_2	0	4167	0
cluster_0	0	0	944

Correct classified: 9,719 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

With Pruning

Cluster \P...	cluster_1	cluster_2	cluster_0
cluster_1	4608	0	0
cluster_2	0	4167	0
cluster_0	0	0	944

Correct classified: 9,719 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

Cross validation using other methods

Logistic Regression

Cluster \P...	cluster_1	cluster_2	cluster_0
cluster_1	2905	1683	20
cluster_2	2268	1894	5
cluster_0	625	317	2

Correct classified: 4,801 Wrong classified: 4,918

Accuracy: 49.398% Error: 50.602%

Cohen's kappa (κ): 0.072%

KNN

K = 7

Cluster \ Cl...	cluster_1	cluster_2	cluster_0
cluster_1	4539	69	0
cluster_2	179	3980	8
cluster_0	0	204	740

Correct classified: 9,259 Wrong classified: 460

Accuracy: 95.267% Error: 4.733%

Cohen's kappa (κ): 0.918%

K = 9

Cluster \ Cl...	cluster_1	cluster_2	cluster_0
cluster_1	4537	71	0
cluster_2	185	3976	6
cluster_0	1	225	718

Correct classified: 9,231 Wrong classified: 488

Accuracy: 94.979% Error: 5.021%

Cohen's kappa (κ): 0.913%

K =19

Cluster \ Cl...	cluster_1	cluster_2	cluster_0
cluster_1	4547	61	0
cluster_2	227	3936	4
cluster_0	0	290	654

Correct classified: 9,137 Wrong classified: 582

Accuracy: 94.012% Error: 5.988%

Cohen's kappa (κ): 0.895%

Random Forest

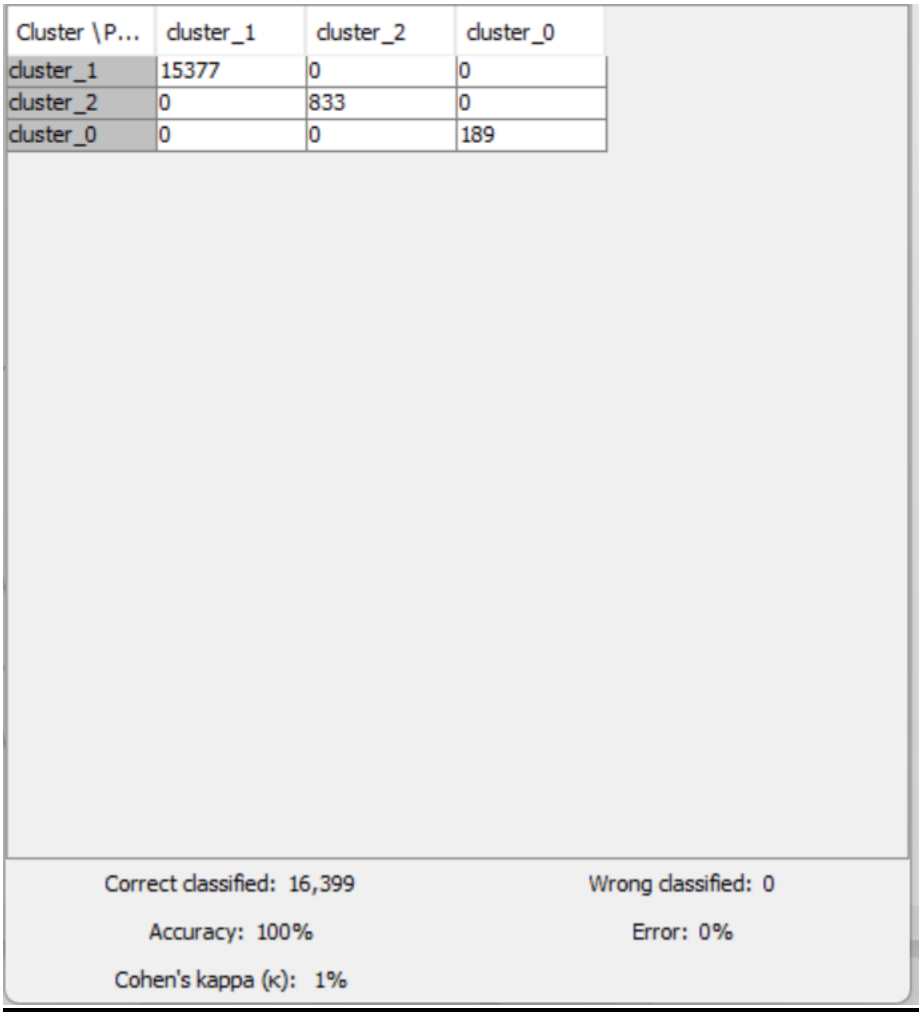
Cluster \P...	cluster_1	cluster_2	cluster_0
cluster_1	851	65	6
cluster_2	141	683	9
cluster_0	112	49	28

Correct classified: 1,562 Wrong classified: 382

Accuracy: 80.35% Error: 19.65%

Cohen's kappa (κ): 0.645%

XGBoost



Cluster \P...	cluster_1	cluster_2	cluster_0
cluster_1	15377	0	0
cluster_2	0	833	0
cluster_0	0	0	189

Correct classified: 16,399 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

Cross validation using Decision Trees: Both with and without pruning show high accuracy of 100% and Cohen's Kappa scores indicating good performance. Pruning has been show to show the same results as without pruning.

• Cross validation using Logistic Regression: This algorithm Shows Low accuracy and Cohen's Kappa score unsimilar to decision trees, indicating less effectiveness of these methods for the dataset.

• Cross validation using KNN: Performs significantly higher compared to other Logistic Regression but lower than Decision Tree. This showed that KNN is better but not better than Decision Tree.

• Random Forest and XGBoost (Ensemble learning): Both ensemble methods perform well with high accuracy and Cohen's Kappa scores. XGBoost outperforms Random Forest slightly in terms of accuracy and Cohen's Kappa, indicating its superior predictive power for this dataset.

For this dataset, ensemble learning methods like Random Forest and XGBoost along with Decision Trees and KNN seem to be the most effective models in terms of accuracy and robustness. Logistic Regression not performs well and provides interpretable results which can be won't be advantageous in certain scenarios.

5. Managerial Insights

Employee compensation data is a valuable asset for managers to understand their workforce, make informed decisions, and promote a fair and competitive work environment. Here are some key managerial insights you can glean from this data:

1. Identify Pay Equity:

- Analyze salary ranges for similar positions, experience levels, and demographics (gender, race, etc.) to identify any potential pay gaps.
- Use metrics like compensation ratios to compare individual salaries to the midpoint of their pay range.
- Investigate and address any pay discrepancies that can't be justified by performance or experience.

2. Benchmarking and Competitiveness:

- Compare your company's compensation packages (salary, benefits) to industry standards and competitors.
- Identify areas where you might be falling behind or exceeding expectations.
- Use this information to make data-driven decisions regarding salary adjustments and benefit offerings to attract and retain top talent.

3. Performance Management and Talent Retention:

- Analyse compensation data alongside performance reviews to identify correlations between pay and performance.
- This can help identify high performers who might be underpaid and at risk of leaving.
- Use compensation adjustments as a tool to incentivize and reward strong performance.

4. Workforce Planning and Budgeting:

- Analyze trends in compensation costs over time to forecast future budgetary needs.
- Identify areas where cost-saving measures might be necessary or strategic investments in talent acquisition are justified.
- Use compensation data to inform workforce planning decisions like promotions, hiring, and restructuring.

5. Employee Satisfaction and Motivation:

- Conduct surveys alongside compensation analysis to understand employee sentiment regarding their pay and benefits.
- Identify areas where compensation might be impacting morale or motivation.
- Use compensation data to create a transparent and competitive compensation strategy that fosters employee satisfaction.

Here are some additional tips for using employee compensation data effectively:

- **Maintain Data Security and Privacy:** Ensure employee compensation data is secure and used responsibly, complying with all data privacy regulations.
- **Transparency and Communication:** Communicate your compensation philosophy and how data is used to employees to build trust and understanding.
- **Regular Analysis:** Conduct regular reviews of compensation data to stay informed about trends and address any emerging issues.

By leveraging employee compensation data effectively, managers can gain valuable insights to create a fair, competitive, and motivating work environment that attracts and retains top talent.